UNIT 11

Statistical pictures

Introduction

In Unit 4, you looked at a variety of statistical summary values that can be used to describe a dataset. Some of these measure location – that is, where the data are centred – these include the mean and median. Others – the range, interquartile range and standard deviation – give an overview of how widely the data are spread.

You will need to use your computer for many of the activities in this unit.

In this unit you will look at a second important set of tools for providing an overview of the data in a dataset. This time, rather than carrying out calculations, you will be focusing on producing and analysing pictures of the data, in the form of statistical graphs and charts, to help to reveal significant features.

Unit 4 also provided you with a number of small datasets. You will be using some of these again, but this time to see how they may be interpreted pictorially.

Another of the big statistical ideas in Unit 4 was the statistical investigation cycle, as shown in Figure 1. You saw that it was useful, in any investigation, to break your thinking down into clear steps, defined by the acronym PCAI (*pose* a question, *collect* the data, *analyse* the data, *interpret* the results).

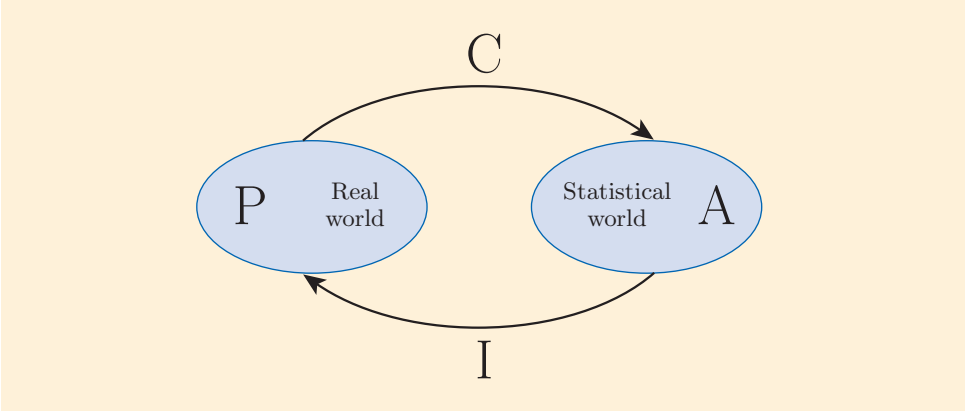


Figure 1 The PCAI statistical investigation cycle described in Unit 4

You will be making use of the PCAI cycle in this unit also, particularly in the final section, where you are asked to apply some of the skills and ways of working of the statistician to investigate a controversial question: Does extra-sensory perception (ESP) exist?

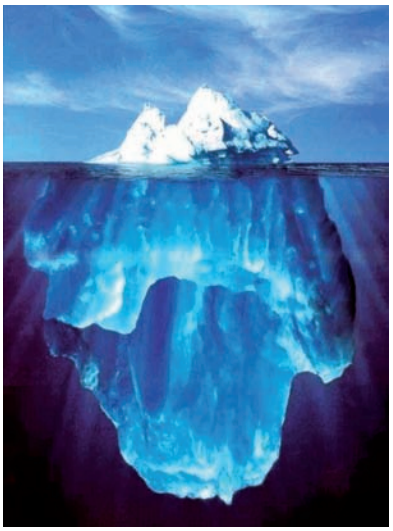


Figure 2 A picture can be worth a thousand words

1 Dotplots and boxplots

In Unit 4 you were asked to investigate the responses of 30 Open University students who were asked to attach a number to how they interpreted the words 'possible' and 'probable', on a scale from 0 to 100 where 0 means impossible and 100 means certain. Any conclusions that you were able to draw were based on summaries of location and spread – in other words, on a purely numerical analysis of the data.

However, it is sometimes said that a picture is worth a thousand words. In this section you will revisit the 'Possible' and 'Probable' data, but now in pictures. You will return to using the computer resource Dataplotter, which can display either one or two datasets in the form of any of four

different types of statistical plot: *dotplots*, *boxplots*, *histograms* or *scatterplots*. (You used Dataplotter in Unit 4 to create dotplots and in Unit 6 to draw scatterplots.)

This first section describes two of these graphical tools: dotplots and boxplots. You will see how these plots are constructed and how they can be used to interpret data. Histograms are introduced in Section 2.

1.1 Dotplots

Dotplots were introduced in Unit 4 as a pictorial way of displaying values in datasets. They are based on the simple principle that each value in a dataset can be represented by a dot positioned above a horizontal axis so that you can see, at a glance, the location and spread of the values.

Dataplotter

More details on how to use Dataplotter are given in Subsection 2.4 of the MU123 Guide.

Activity 1 *Using Dataplotter to draw a dotplot*

Open Dataplotter. There may still be data from previous activities in one or both of the two data columns on the left. If so, click the 'New' button under each column. Then click the 'Dotplot' tab, if it is not already selected.

(a) Position the cursor in cell 1 of the left-hand data column, click, and enter the numbers 1, 2, 5, 8, 15, pressing Enter after each entry (including the last). Watch what happens in the graphing area as each number is entered. Notice that the horizontal axis automatically rescales to accommodate each new number.

(b) Spend a few moments making sure that you understand what the ten summary values on the right of the software display mean. For example, the final summary value, labelled 'n', tells you how many data values there are in the dataset.

(c) Based on your work in Unit 4, explain in your own words the meaning of the summary values median and standard deviation.

The dataset that you have created in this activity will not be needed again, so you may wish to delete it. To do this, click the 'Delete dataset' button at the bottom of the column.

In the next activity, you are asked to compare the dotplots for the 'Possible' and 'Probable' datasets. The dotplots provide a useful overview of the data, allowing you to see where each value is positioned in relation to the others. Values that are repeated are stacked up in vertical lines, which makes it easy to see at a glance how many students gave a particular value.

Dataplotter

Activity 2 *Comparing dotplots*

(a) In Dataplotter, click on the drop-down menu for the left-hand column and select the dataset '# Possible'. Click on the 'Dotplot' tab if it is not already selected.

How many students gave the value 30?

(b) Now click on the drop-down menu for the right-hand column and select the dataset '# Probable'. You should now see the two dotplots corresponding to the '# Possible' and '# Probable' datasets, one below the other. Based on these plots, identify two basic differences between these datasets.

1.2 Boxplots

As you have seen, a dotplot is a simple but powerful form of graphical representation, where every value in a dataset is represented by a dot above a horizontal axis. However, it is not always sensible to include every single value in a picture representing a dataset, particularly if the dataset is large. Instead, a graphical representation showing a selection of some key values from the dataset can be used.

The boxplot was invented by the American statistician John Tukey (1915–2000). Tukey was also partly responsible for the word 'software', referring to computer programs as opposed to the machines that they run on (the hardware).

A **boxplot** is a popular way of depicting data, based on showing the locations of the following five key summary values of a dataset:

minimum value (Min),
lower quartile (Q1),
median (also known as Q2),
upper quartile (Q3),
maximum value (Max).

You should already be familiar with these five summary values from Unit 4, but the next paragraph gives a brief reminder of what they are and how to find them.

For a given dataset, first sort the data into ascending order. Look at the beginning and end of the sorted dataset for the minimum and maximum data values. To find the median, look at the middle of the sorted dataset. If there is an odd number of data values, the median is the middle data value. If there is an even number of data values, the median is the mean of the two middle data values. To find the lower and upper quartiles, look at the data values that are, respectively, one-quarter and three-quarters of the way through the data. The lower quartile is the median of the lower half of the dataset, and the upper quartile is the median of the upper half of the dataset (with the middle data value in the dataset thrown out if the number of data values in the dataset is odd).

For example, here is the 'Possible' dataset, in ascending order, with the positions of these five summary values marked.

1 1 1 1 5 10 10 20 20 30 30 30 30 30 30 35 40 50 50 50 50 50 50 50 60 70 80 85 90 98

↑ Min ↑ Q1 ↑ Median ↑ Q3 ↑ Max

The summary values for this dataset are

$$\text{Min} = 1, \quad \text{Q1} = 20, \quad \text{Median} = 32.5, \quad \text{Q3} = 50, \quad \text{Max} = 98.$$

The median is $(30 + 35)/2 = 32.5$.

For large datasets, you may prefer to use Dataplotter to calculate the summary values.

In this module, you can either draw boxplots by hand or you can use Dataplotter.

Drawing boxplots by hand

The steps below explain how to draw a boxplot by hand. It is best to use squared paper or graph paper.

Boxplots can be drawn either horizontally or vertically, but in this module we will draw them horizontally.

Step 1 Draw a horizontal axis and mark a scale that covers values from the minimum data value to the maximum data value. Add an axis label, including units if appropriate. Also add a suitable title, and the source of the data.

Figure 3 shows this step for the 'Possible' dataset. Since the data values are just numbers, no units are given in this case.

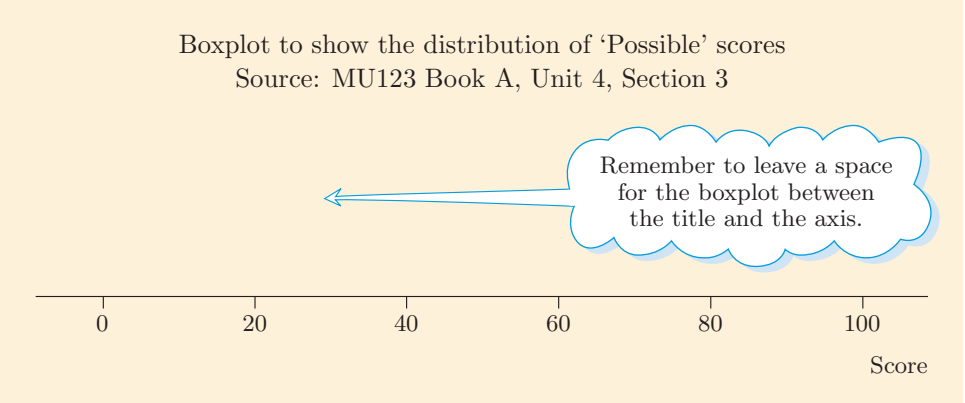


Figure 3 The labelled axis, title and source for the 'Possible' dataset

Step 2 Mark the locations of the minimum, lower quartile, median, upper quartile and maximum, as five vertical lines above the axis. Usually shorter vertical lines are used for the minimum and maximum than for the other three values.

Figure 4 shows this step for the 'Possible' dataset.

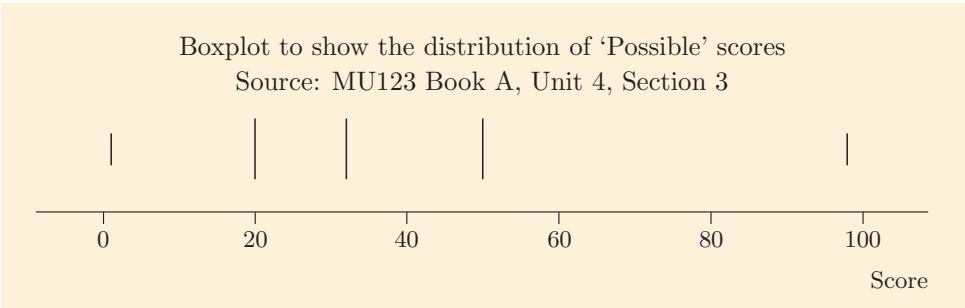


Figure 4 The locations of the five key summary values for the 'Possible' dataset

Step 3 Draw a box around the lower and upper quartiles (to include the median) and draw two lines (called **whiskers**) between the box and the minimum and maximum values. (These features are included simply to make the picture clearer.) Finally, write the five key summary values on the boxplot, so that it is clear at a glance what these values are.

Figure 5 shows this step for the 'Possible' dataset.

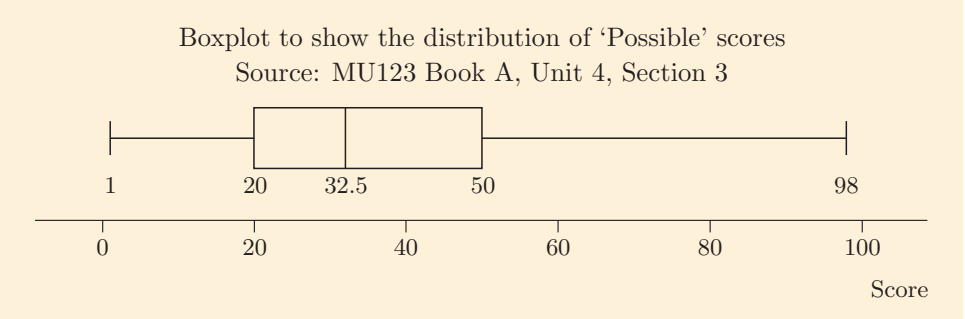


Figure 5 The completed boxplot for the 'Possible' dataset

Figure 6 An entrant from the World Beard and Moustache Championships 2007 held in Brighton, England

Notice the characteristic appearance of a boxplot, which is a central box with horizontal lines to either side that look rather like a pair of waxed whiskers, such as those in Figure 6. In fact, an alternative name for a boxplot is a *box-and-whiskers diagram*. The box is drawn between the quartiles; the whiskers reach out to the extremes of the dataset.

The method for drawing a boxplot is summarised below.

How to draw a boxplot

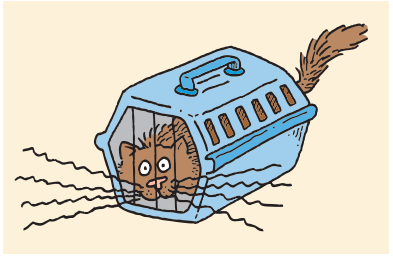
Step 1 Draw the axis, mark the scale and add an axis label, including units if appropriate. Add a title and the source of the data.

Step 2 Mark the five key summary values Min, Q1, Median, Q3, and Max with vertical lines.

Step 3 Add the box and whiskers and label the vertical lines with the five key summary values.

It is worth spending a few moments thinking about the information that can be gleaned from a boxplot. For example, as shown in Figure 7, the length of the entire boxplot represents the *range* of the dataset – the difference between the minimum and maximum values. The length of the box component represents the *interquartile range* (IQR) – the difference between the lower and upper quartiles.

The range and interquartile range of a dataset were introduced in Unit 4.



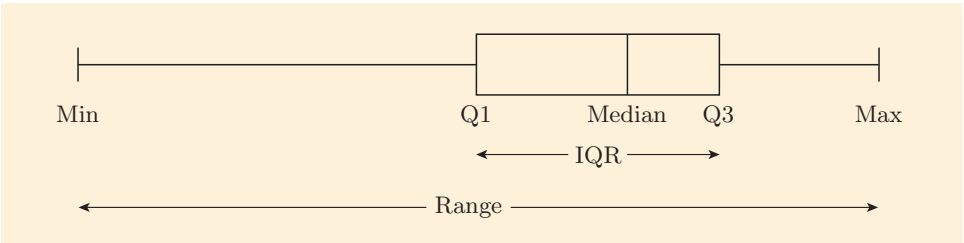


Figure 7 The range and interquartile range as seen on a boxplot

The main features to take note of in a boxplot are the five vertical lines that mark the locations of the five key summary values (Min, Q1, Median, Q3 and Max). Everything else (the box and the two horizontal whiskers) is included to identify which summary value is which.

A special feature of boxplots is that they enable you, quickly and easily, to make remarks about, say, 'the upper quarter' of the data, or 'the middle 50%' of the data, and so on. So a boxplot provides a simple and useful summary of some of the key features of a dataset.

Drawing boxplots on Dataplotter

The boxplots produced by Dataplotter are the same as the hand-drawn ones described above, except that the five key summary values are listed in the right-hand column rather than being displayed on the boxplot itself.

Activity 3 illustrates how two datasets can be compared by using Dataplotter to draw their boxplots one below the other, using the same axis.

Activity 3 *Using Dataplotter to draw boxplots*

Dataplotter

In Dataplotter, select the '# Possible' and '# Probable' datasets, if they are not already loaded. Then select the 'Boxplot' tab to display the boxplots, one above the other.

(a) As you did with the dotplot in Activity 1, spend a few moments relating the boxplots to the summary values given. For example, confirm that the five key positions on the 'Possible' boxplot match up with the five key summary values Min, Q1, Median, Q3 and Max.

(b) Based on the second boxplot on your screen, write down the following values for the '# Probable' dataset:

(i) the median

(ii) the upper quartile

(iii) the lower quartile.

(c) Which of the following statements are correct?

(i) All the 'Probable' data are greater than the upper quartile of the 'Possible' data.

(ii) The interquartile range of the 'Probable' data is greater than that of the 'Possible' data.

(iii) The bottom half of the 'Possible' data is more widely spread than the bottom half of the 'Probable' data.

(d) Comment on the differences between the two boxplots, with particular reference to the five key summary values. How can these differences be interpreted?

As with other statistical charts, if you wish to use a Dataplotter boxplot to present results from a statistical investigation or as part of an assignment question, then remember to include a title and the source of the data. Figure 8 shows a title and source written above a boxplot, but you can also write these below the chart, or anywhere that seems appropriate. For a boxplot, you should also include the list of summary values, as shown in Figure 8.

There are instructions for producing a screenshot of Dataplotter or Graphplotter in Subsection 2.4 of the MU123 Guide.

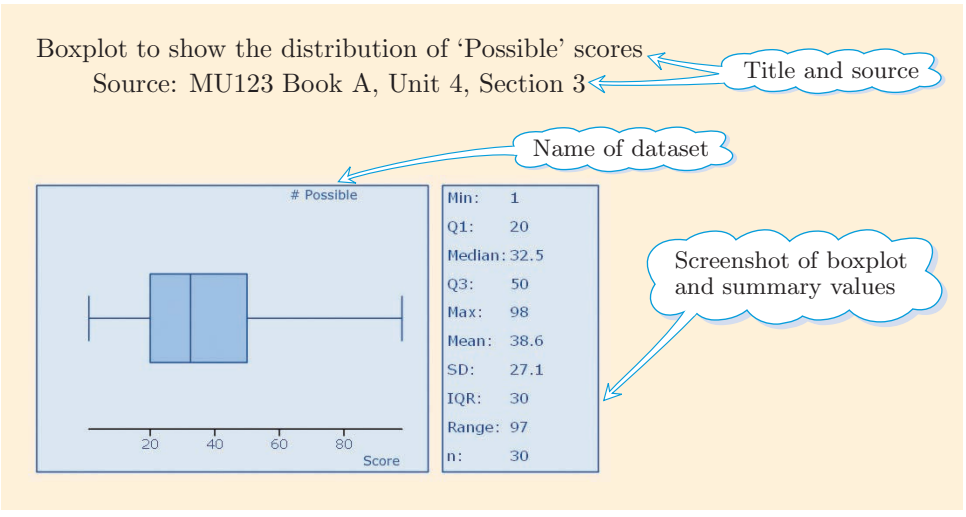


Figure 8 Presenting a Dataplotter chart

Shapes of boxplots

In the activity below, you are asked to type some simple numbers into one of the data columns in Dataplotter, to find out more about what the shape of a boxplot can reveal about the data that it represents.

Dataplotter

Activity 4 *Investigating the shapes of boxplots*

Use Dataplotter, with the 'Boxplot' tab selected.

(a) Click each of the two buttons marked 'New' to open a new, blank dataset in each of the two columns. Then enter the following numbers into the left-hand column: 1, 2, 3, 4, 5, 6, 7.

As the numbers in this dataset are symmetrically spaced, you may not be surprised that the corresponding boxplot is also symmetrical.

Again, try to match up the key features of the boxplot with the five corresponding summary values listed on the right of the screen.

(b) Now change the 7 to 20 (by selecting the seventh cell and overtyping with 20 followed by Enter), and observe the effect on the shape of the boxplot. Check the new scale carefully. Have any of the five key summary values (Min, Q1, Median, Q3, Max) changed? Can you explain why some values have not changed?

(c) Has the mean changed? If so, can you explain why?

An important feature of the five key summary values depicted in a boxplot is that, provided that the dataset contains six or more values, making either of the two extreme values (the minimum or maximum) more extreme will have no effect on the other four summary values.

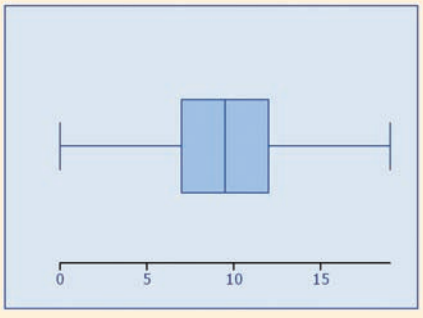
Activity 5 *Matching datasets and boxplots*

(a) Look at these two datasets (both ordered from smallest to largest).

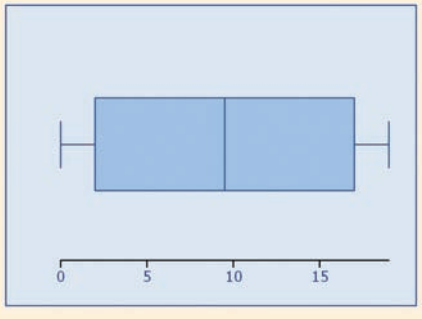
Dataset 1	0	6	7	8	9	10	11	12	13	19
Dataset 2	0	1	2	3	9	10	16	17	18	19

Try to match each dataset to its corresponding boxplot below.

Boxplot A

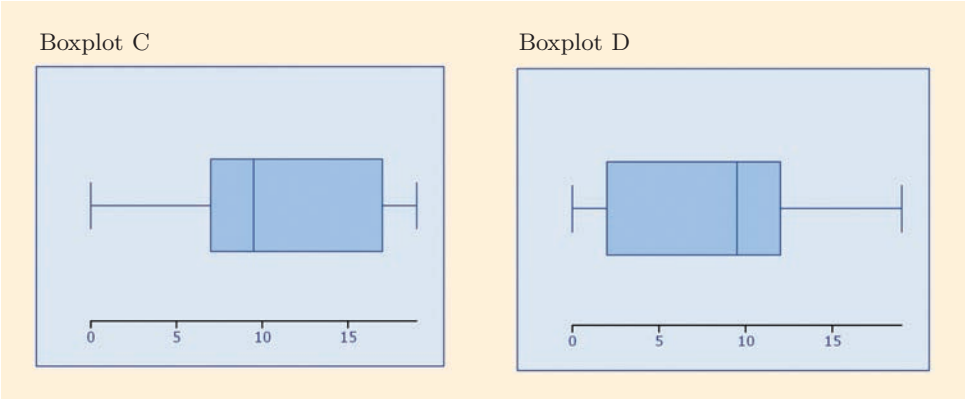


Boxplot B



(b) Now look at these two datasets and try to match each one to its corresponding boxplot below.

Dataset 3	0	1	2	3	9	10	11	12	13	19
Dataset 4	0	6	7	8	9	10	16	17	18	19



There are several general features of the shape of a boxplot that are worth thinking through. Each of the four sections of a boxplot – the two parts of the box and the two whiskers – represents approximately the *same number* of data values – a quarter of the total number. For example, the left-hand whisker represents all the data values lying between the minimum value and the lower quartile, and the left-hand part of the box represents all the data values lying between the lower quartile and the median. So if one of these four sections of a boxplot is comparatively long, then the data in the corresponding quarter of the dataset are comparatively spread out – in other words, the data values are less densely packed together.

For example, Boxplot C in Activity 5 has a long left whisker and a short right whisker. This reflects the fact that the data in the bottom quarter of Dataset 4 are more spread out than those in the top quarter. Similarly, the left part of the box in Boxplot D is longer than the right part, which reflects the fact that the data in the second quarter of Dataset 3 are more spread out than those in the third quarter.

So a boxplot provides a useful picture of how the data are spread.

In the boxplot in Figure 9(a), the left whisker is longer than the right whisker, and the section of the box to the left of the median is longer than the section of the box to the right of the median. This indicates that the smaller values in the dataset are more spread out than the larger values. In cases like this, the dataset and the boxplot are said to be **left-skewed**. The boxplot in Figure 9(b) indicates that the larger values in the dataset are more spread out than the smaller values. A dataset like this, and its boxplot, are said to be **right-skewed**.

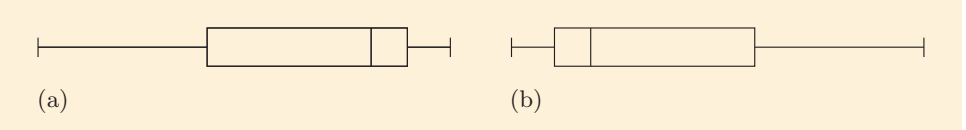


Figure 9 (a) A left-skewed boxplot (b) A right-skewed boxplot

If all four regions of a boxplot are approximately equal in length, then the data tend to be fairly uniformly spread between the minimum and maximum values. However, an interesting phenomenon about data drawn

Measurements like those listed in the paragraph here often concentrate in the middle in a characteristic way, which is described by a famous mathematical function called the *normal distribution* or *Gaussian distribution*. The second name refers to the great mathematician Karl Friedrich Gauss, who was mentioned in Units 3, 7 and 9.

Dataplotter

from the natural world (for example, people's heights, weights of robin eggs, midday temperatures at a particular location in June, lengths of human gestations, and so on) is that they tend not to be uniformly spread. Instead there is, typically, a concentration of values in the middle, with values spread more sparsely at the extremes. As you'll see from the next activity, this means that such data tend to produce boxplots with whiskers that are longer than either of the two parts of the box.

Activity 6 *Using a boxplot to investigate the spread of data*

Use Dataplotter, with the 'Boxplot' tab selected.

Select the dataset '# Weight baby' (not the scatterplot version '# SP Weight baby' that you used in Unit 6) from the left-hand drop-down list. This dataset contains the weights, in kg, of 32 babies, which were listed in a table in Unit 4.

If there are any data in the right-hand column, then remove them to avoid distractions. You can do this by clicking 'New' to open a new, blank dataset. Alternatively, if you created the data in the right-hand column but no longer want them, then you can click 'Clear' to remove them from the dataset.

(a) Look at the boxplot that represents the baby weights. Which are longer, the whiskers or the two parts of the box?

(b) What does this tell you about how the data are spread?

The data in some other types of datasets also tend to be spread in a similar way to the data from nature discussed above – bunched in the middle and more sparse elsewhere. For example, when a quantity is measured there is a small difference between the true value of the quantity and the measurement – this is known as a *measurement error*, and it can be either positive or negative, depending on whether the measurement is larger or smaller than the true value. It is important to consider measurement errors in subjects like physics. Measurement errors tend to bunch around zero, with very high and very low errors less common, so boxplots of measurement errors tend to have narrow boxes and long whiskers.

To summarise this subsection, the following box presents three key facts to remember when studying the shape of a boxplot.

Characteristics of boxplots

- A boxplot is composed of four sections (two whiskers at either end and two sections within the central box), each of which contains approximately the same number of data values.
- Where a particular section of a boxplot is narrow, this indicates a dense concentration of the data, whereas a wide section indicates where the data are more sparsely spread.
- A boxplot with a narrow box and long whiskers indicates that the data are concentrated in the middle and more widely spread at the extremes. This is typical of data drawn from nature.

1.3 Investigating poverty levels

According to Help the Aged's report *Spotlight on Older People in the UK*, which was published in January 2009, one in five older people lives 'in poverty'.

Clearly most people have a general sense of what it means to be in poverty, but when it comes to serious debate about the issue, such labels need to be tied down more quantitatively, to ensure that everyone is referring to the same phenomenon. There are several different definitions of poverty, used by different institutions. The World Bank defines poverty in *absolute* terms as living on less than \$1.25 per day (at the time of writing). However, this is not a suitable definition for richer countries such as those in the European Union, where a *relative* figure is used.

Absolute and *relative* comparisons were explained in Unit 1.

Read the extract below, which includes a commonly-used definition of the term *poverty* in the UK and other European countries.

> The widely accepted definition of poverty is having an income which is less than 60% of the national average (excluding the wealthiest members of society). On this measure, the proportion of the UK population defined as in poverty is roughly one in five. And this roughly one in five figure has remained stubbornly high through both Conservative and Labour governments.
>
> Julian Knight, 'The changing face of poverty', BBC News, 26 July 2005

A more precise formulation of the definition of poverty referred to in the extract is that a person is in poverty if they live in a household with an income that is less than 60% of the national median household income. The amount of money that is 60% of the national median household income is referred to as the *poverty threshold*.

Sometimes adjustments are made to this definition of poverty to make it appropriate for households of different sizes, for example. But, to keep things simple, we will work with this basic definition, and we will look at households each containing just one person, who is an earner.

You may be wondering why the one-in-five figure mentioned in the extract does remain so 'stubbornly high', even as earnings change over time. In the next activity boxplots are used to illustrate the changes in earnings as different types of pay rises are applied, and you are asked to investigate the effect of these pay rises on the number of people in poverty.

Activity 7 *Investigating different types of pay rise*

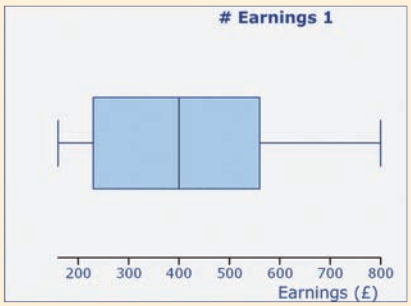
Dataplotter

Use Dataplotter, with the 'Boxplot' tab selected.

In the left-hand data column select the dataset '# Earnings 1', which consists of the hypothetical weekly earnings of 13 people, each of whom is the only person in his or her household. These are shown below.

> Hypothetical weekly earnings of 13 people (£)
>
> 160, 190, 220, 240, 290, 350, 400, 480, 510, 530, 590, 600, 800.

The corresponding boxplot should look like this.



Now consider the poverty threshold in the microworld of just these 13 earners. Suppose that it is defined to be 60% of the median earnings, as discussed above.

(a) Calculate the amount of money that is the poverty threshold.

Hence calculate the number of people and the percentage of people in poverty.

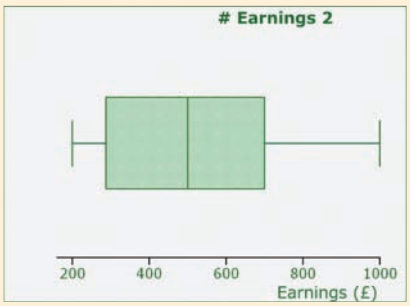
(b) Now let's increase everyone's wages by 25%.

Revised weekly earnings of the 13 people (£)

200, 237.5, 275, 300, 362.5, 437.5, 500, 600, 637.5, 662.5, 737.5, 750, 1000.

In the right-hand data column, open the dataset '# Earnings 2', which contains the revised earnings of these 13 people, based on an across-the-board percentage increase of 25%.

The corresponding boxplot is shown below.



Based on these revised earnings, calculate the new poverty threshold.

How has this pay rise affected the number and percentage of people in poverty?

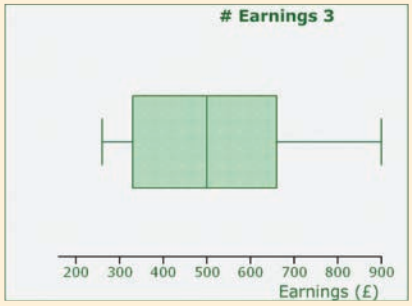
(c) Now, rather than increasing everyone's earnings by the same percentage, let's give everyone a fixed pay rise of £100.

Revised weekly earnings of the 13 people (£)

260, 290, 320, 340, 390, 450, 500, 580, 610, 630, 690, 700, 900.

Again in the right-hand data column, open the dataset '# Earnings 3', which contains the revised earnings of these 13 people, based on an across-the-board £100 increase.

The new boxplot is shown below.



Based on these revised earnings, calculate the new poverty threshold.

How has this pay rise affected the number and percentage of people in poverty?

To understand the reasons behind the results obtained in Activity 7, consider a group of earners (of any size), with each earner the only person in his or her household. Suppose that the median of the earners' weekly wages, in £, is m, and suppose, as in Activity 7, that the poverty threshold is based just on this group of earners. Then the poverty threshold, in £, is $0.6m$.

Now suppose that everyone is given a 25% pay rise. Then everyone's wage is 1.25 times what it was before, so the new median wage is also 1.25 times what it was before. That is, the new median wage, in £, is

$$1.25m$$

and hence the new poverty threshold, in £, is

$$0.6 \times 1.25m,$$

which is the same as

$$1.25 \times 0.6m.$$

This is 1.25 times the old poverty threshold. So the poverty threshold has risen by 25%. As everyone's wage has also risen by 25%, no one has moved over the poverty threshold.

Now suppose, instead, that everyone is given a pay rise of £100. Then the new median wage is also £100 more than it was before, so the new poverty threshold, in pounds, is

$$0.6 \times (m + 100).$$

Multiplying out the brackets gives

$$0.6m + 60.$$

This is the old poverty threshold, plus 60. So the poverty threshold has risen by £60. But everyone's wage has risen by £100, so anyone whose wage was £40 or less below the poverty threshold, before the wage rise, has now moved over the poverty threshold and is no longer in poverty.

Activity 7 raises some important questions about wage increases and their effect on income inequalities. A particular question of interest is: How can wage increases be organised so as to reduce the number of people in poverty?

A fixed percentage increase means that, in absolute terms, the high wage earners get a bigger rise than the lower earners, thereby stretching the boxplot wider. This means that absolute inequalities widen. Relative inequalities are maintained, and there is no change in the number of people whose earnings lie below the poverty threshold.

With a wage increase of a fixed amount, however, a different picture emerges. Here, the boxplot is not stretched but simply moved to the right by the amount of the rise (in the case of Activity 7, by £100). So in absolute terms, the amounts of the inequalities are maintained (in Activity 7 the highest earner earns £640 more than the lowest earner both before and after the rise). However, in relative terms, such an arrangement is much better for the lowest earners, as for them a £100 rise represents a larger percentage increase than is the case for the most well-off. This type of pay rise is one way to reduce the number of people below the poverty threshold.

In general, if poverty is measured in the way discussed in this subsection, then the way to achieve a reduction in the number of people below the poverty threshold is to increase the household incomes of the low-income households by a greater percentage than those of the middle-income households. This can be achieved in various ways, such as by awarding greater percentage wage increases to low earners than to middle earners, increasing the number of earners in low-income households, reducing the amount of tax that low earners pay or increasing the benefits paid to low-income households.

In this section, you have seen how to use dotplots and boxplots to illustrate datasets, and how to interpret these plots. Dotplots are useful for small datasets, as they give a quick impression of all the data values. Boxplots tend to be more useful for large datasets, as they summarise data by displaying key summary values, although the detail of the individual data values is lost.

2 Histograms and bar charts

2.1 Histograms

In the last section you saw how to use dotplots and boxplots to provide pictorial representations of datasets.

A third way to represent a dataset pictorially is to use a *histogram*. For example, consider the dataset below, which consists of the weights in kilograms of 32 women at the start of their pregnancies. It is part of the backache dataset that was given in Unit 4.

54.5 59.1 73.2 41.4 55.5 70.5 76.4 70 52.3 83.2 64.5 49.5
70 52.3 68.2 47.3 66.8 70 56.4 53.6 59.1 44.5 55.9 57.3 69.5
73.2 62.7 73.6 70 56.7 58.2 68.2

This dataset can be represented by the chart in Figure 10. The chart shows that there are two data values in the interval 40–45, two data values in the interval 45–50, four data values in the interval 50–55, and so on.

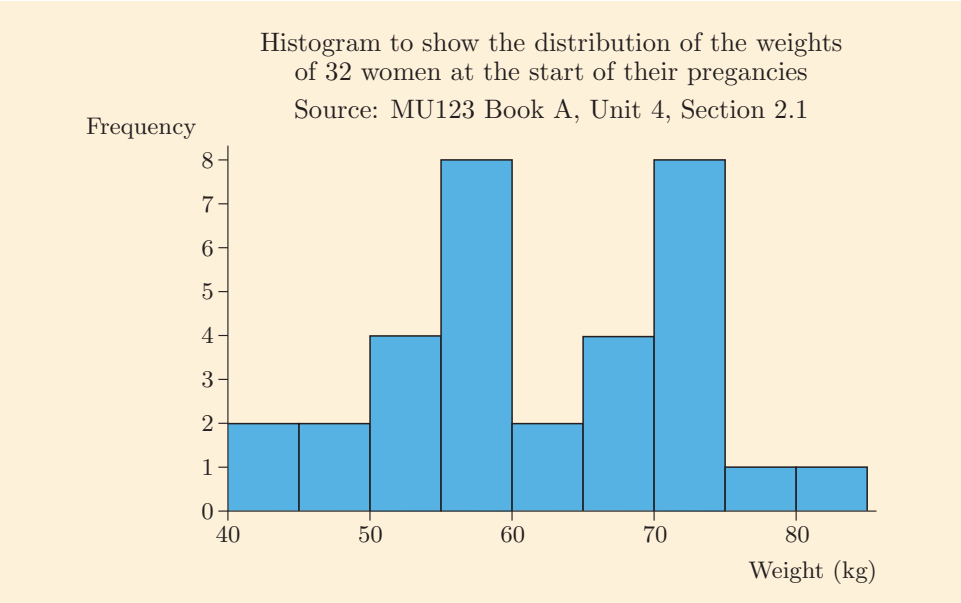


Figure 10 A histogram representing the dataset of women's weights

Charts like this are called **histograms**. In general, if you want to draw a histogram to illustrate a dataset, then the first step is to choose a sequence of contiguous intervals that together cover all the values in the dataset. The intervals are usually chosen to have equal widths, and, if this is the case, then for each interval you just draw a rectangle whose height is proportional to the number of data values that lie in that interval, in the way shown in Figure 10. These rectangles are called *bars* or sometimes *columns*.

Contiguous means 'touching'.

A data value that lies on the boundary between two intervals is usually included in the bar to the *right* of the boundary. There is no obvious reason for this choice – it is just the usual convention. So, for example, in the histogram in Figure 10, the bar covering the interval 70–75 represents values that are *at least* 70 but *less than* 75, and the three values of 70 kg in the dataset are represented in this bar. In general, each interval includes the left boundary value but not the right boundary value.

The left boundary value of the *first* interval is called the *start value* of the histogram. For example, the start value of the histogram in Figure 10 is 40.

The vertical scale of the histogram in Figure 10 indicates the frequencies with which the data values in the different intervals occur, but other types of vertical scale can be used, provided that the heights of the bars are proportional to the frequencies of the data values in the intervals. For example, the vertical scale in Figure 10 could instead have indicated the *percentages* of the women in the survey whose weights fell into the various intervals. In that case, it would have gone from 0% to 25% (because the highest bars, indicating frequencies of 8, each correspond to 25% of the 32 women).

A histogram whose vertical scale shows frequencies (rather than percentages, for example) is sometimes called a frequency diagram. In general, a *frequency diagram* is any diagram that shows the frequencies of particular items, values or groups of values.

The shape of a histogram indicates the distribution of the data values in the dataset that it represents. For example, in the histogram in Figure 10, the heights of the bars are fairly similar, apart from two peaks at around 55 kg and 70 kg. A histogram can show more detail than a boxplot about the distribution of the data values in a dataset. As you have seen, a boxplot just shows five summary values.

You learned about the distinction between discrete and continuous data in Unit 4.

Histograms are usually used to represent *continuous* data, such as measurements. However, if the data values in a dataset are discrete but take a large number of different values, then they are often treated as continuous and can be represented by a histogram. For example, you could draw a histogram of examination scores, which may take any whole-number value from 0 to 100.

2.2 Bar charts

If the data values in a dataset are discrete and take only a small number of values, then it is usual to represent them not by a histogram, but by a similar-looking type of statistical chart called a **bar chart**. When a dataset is depicted by a bar chart, each bar corresponds to a *single* possible data value, rather than an interval of possible data values. The bars in a bar chart are drawn with gaps between them to reflect this fact.

For example, the bar chart in Figure 11 shows the frequencies with which the six scores 1, 2, 3, 4, 5, 6 cropped up as a result of 30 rolls of a die.

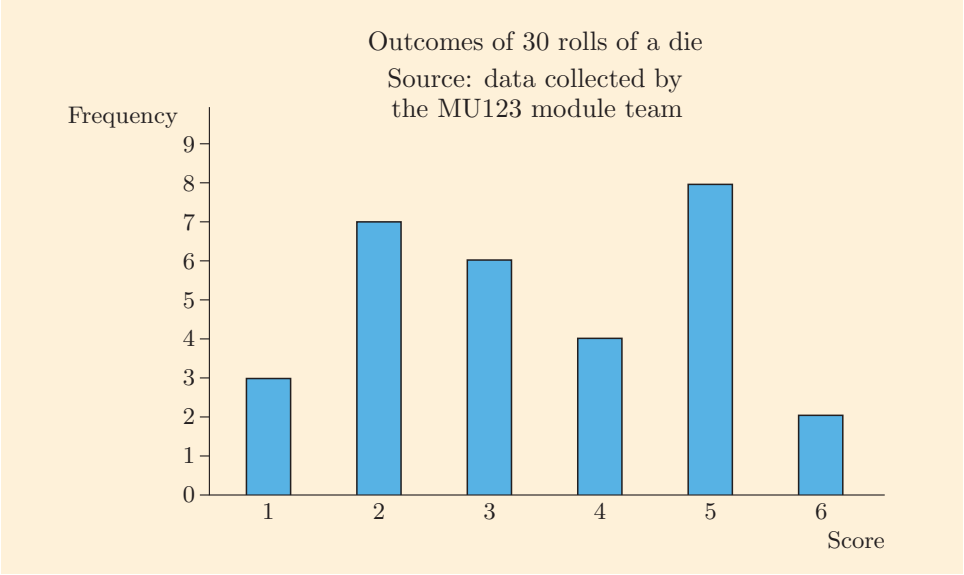


Figure 11 A bar chart

The data values represented by a bar chart do not have to be numbers – they can be 'categories', such as colours, countries, and so on. Data of this type are called **categorical data**. For example, Figure 12 shows the various types of housing accommodation used in England in a particular year (2007–08). The length of each bar corresponds to the percentage of all accommodation that was of that particular type. Because the categories are non-overlapping and cover *all* accommodation in England, the percentages add to 100.

The bars in a bar chart can be drawn either vertically or horizontally, as you can see from Figures 11 and 12. In contrast, the bars in a histogram are nearly always drawn vertically.

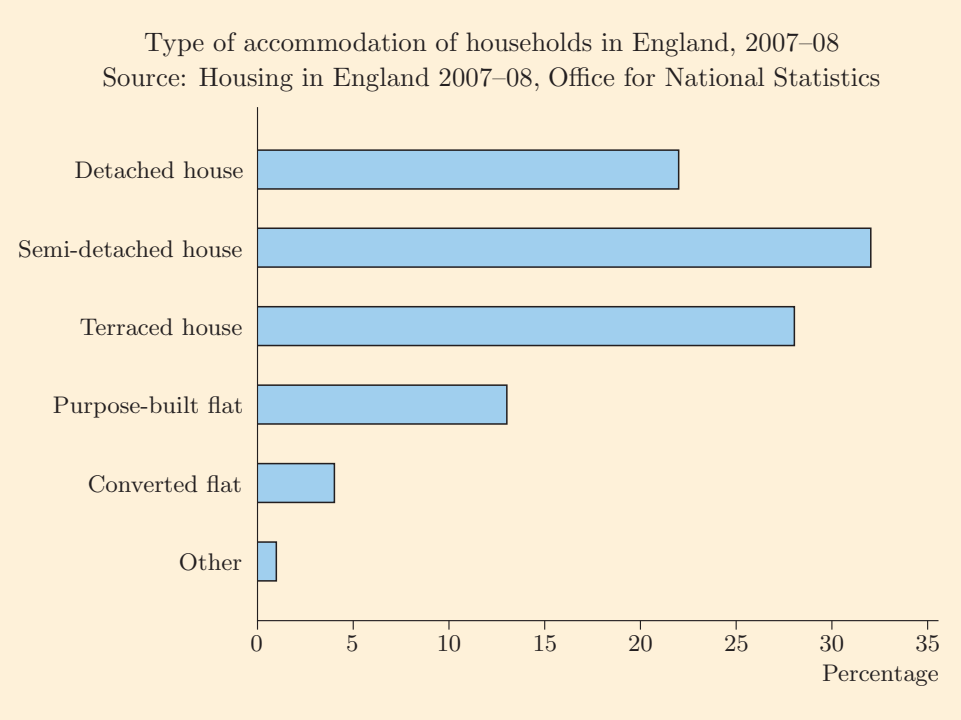


Figure 12 A bar chart representing categorical data

2.3 Exploring histograms

In this subsection you will use Dataplotter to explore histograms. The histograms drawn by Dataplotter have equal interval widths, and they have frequency of occurrence on the vertical axis.

In the next activity you will use Dataplotter to explore different histograms for the dataset consisting of the weights of 32 women at the start of their pregnancies that was given on page 20. One possible histogram is given in Figure 10 on page 21, but you can produce different histograms by changing the start value and/or interval width.

Activity 8 *Exploring histograms with Dataplotter*

Dataplotter

Use Dataplotter, with the 'Histogram' tab selected.

(a) Select the dataset '# Weight start' (not '# SP Weight start') in the first data column. Remove any data in the second column – you can do this by clicking 'New' to create a new, blank dataset.

Locate the 'Start value', 'Interval' and 'Auto' boxes just below the histogram. When 'Auto' is ticked, as it is by default, the software chooses a start value and an interval width for the histogram automatically. Click in the box to switch 'Auto' off, then set 'Start value' to 40 and 'Interval' (which means interval width) to 5.

You should see the same histogram as in Figure 10 on page 21. How many data values are there in the interval 65–70?

(b) The histogram has two peaks at around 55 and 70. Now try increasing the interval width from 5 to 10 and note where the peaks occur.

Then reduce the interval width to 2 and observe the effect.

Finally, reduce the interval width to 1 and observe the histogram again.

In general, what effect do you find that the interval width has on the shape of the histogram?

In Activity 8, you saw that a small interval width results in a 'spiky' histogram that displays lots of detailed features in the distribution of the data but may fail to give a good picture of its overall shape. On the other hand, a large interval width results in a 'lumpy' histogram that displays the large-scale structure in the distribution of the data, but at the expense of the detail. A good choice of interval width is usually a compromise between these two extremes.

The 'Auto' setting gives reasonable choices for the start value and interval width and these are often adequate. However, in order to create a histogram exactly as you want it to appear, you may need to choose these values yourself.

Sometimes you may need different interval widths in different parts of the horizontal axis in order to produce a good overall picture. This happens when there are many data values in some parts but few in others. In cases like these, different interval widths can be used, but then the number of data values in an interval is represented not by the height of the corresponding bar, but by its *area*.

For example, suppose that a statistical investigation has been carried out to find the weekly earnings of a large group of people, and the results are illustrated by the histogram in Figure 13. Notice that the intervals on the left are narrow and widen as you go to the right, where earnings are higher. The reason for drawing the histogram like this is that the narrow intervals for the lower earnings show the shape of the distribution well, but there are relatively few people with the higher earnings and so it makes sense to have wider intervals here, to avoid spikiness. The size of each column must now be judged on the basis of the key alongside the chart, which indicates the size of the area corresponding to 500 people.

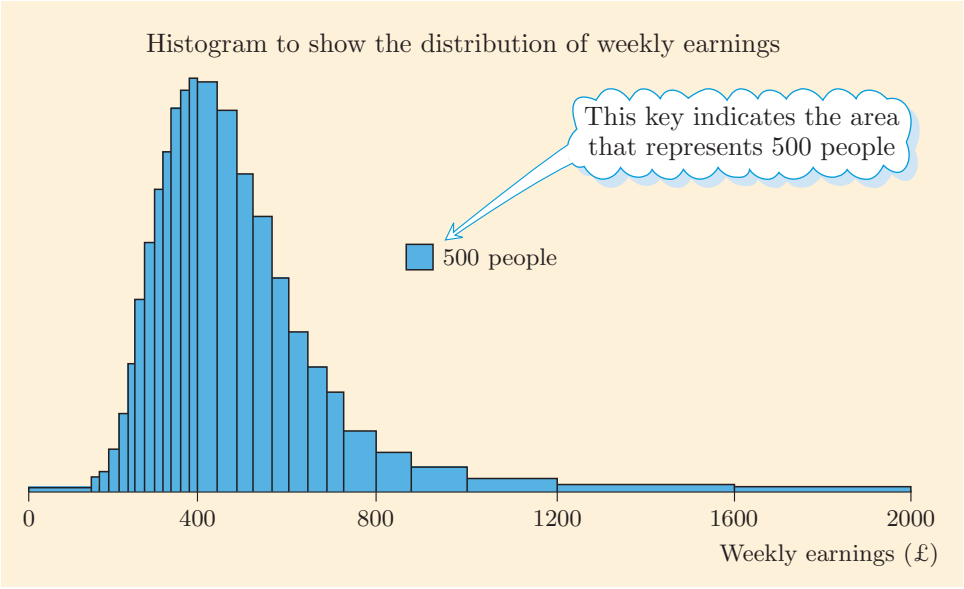


Figure 13 A histogram with unequal interval widths

Histograms with unequal interval widths are relatively uncommon, and all the histograms in the remainder of the module have equal interval widths. As mentioned earlier, Dataplotter can plot only histograms with equal interval widths.

There is another thing worth noticing about the histogram in Figure 13: it is asymmetric, with its right tail longer than its left tail. This indicates that the dataset is right-skewed, and the histogram itself is also described as being right-skewed. Similarly, a histogram whose left tail is longer than its right tail indicates that the dataset is left-skewed, and is itself said to be left-skewed.

In the next activity, you are asked to use histograms to compare the data in two datasets.

Dataplotter

Activity 9 *Using histograms to compare datasets*

Use Dataplotter, with the 'Histogram' tab selected.

(a) Make sure that '# Weight start' dataset is still in the first data column, and select the '# Weight end' dataset in the second column.

Make sure that the start value and interval width for the first histogram are set to 40 and 5, respectively. Uncheck the 'Auto' box for the second histogram, and set its start value and interval width to the same values.

What does the 'Weight end' histogram tell you about the distribution of the weights of the women at the end of their pregnancies?

(b) Compare the shape and position of the 'Weight end' histogram with those of the 'Weight start' histogram. How do you account for the differences between them?

(c) Now select the 'Boxplot' tab, and describe what the boxplots show. Which of these two types of chart do you think is easier to use if you want to compare the two datasets?

In the next activity you are asked to match up some small datasets with histograms that represent them.

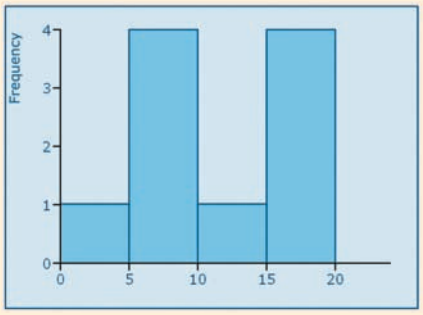
Activity 10 *Matching datasets and histograms*

In Activity 5 you were asked to match up each of four datasets consisting of ten integer values with their corresponding boxplots. These datasets are reproduced below.

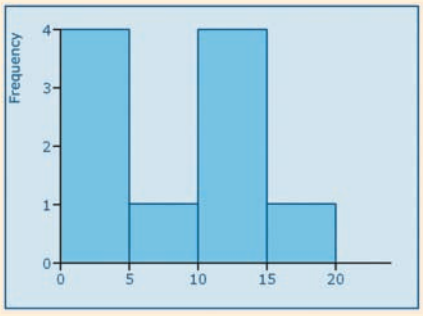
Dataset 1	0	6	7	8	9	10	11	12	13	19
Dataset 2	0	1	2	3	9	10	16	17	18	19
Dataset 3	0	1	2	3	9	10	11	12	13	19
Dataset 4	0	6	7	8	9	10	16	17	18	19

(a) The figure overleaf shows Dataplotter histograms representing these four datasets. Without using Dataplotter, work out which histogram corresponds to which dataset. All four histograms have been set to have an interval width of 5.

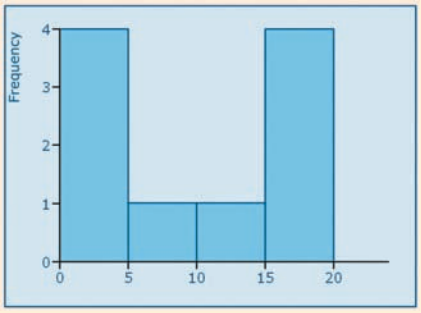
Histogram A



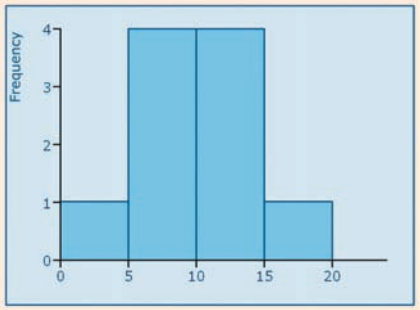
Histogram B



Histogram C



Histogram D



(b) Now compare the histograms in part (a) with the boxplots of the same datasets in Activity 5 on pages 14–15. How are the shapes of the boxplots reflected in the shapes of the histograms?

Histograms and boxplots both provide useful pictures of data. They each have their advantages and disadvantages.

Boxplots give a quick overall impression of how the data in a dataset are distributed, and important statistical summary values for the data can be read directly off them. They are also good for comparing datasets.

Histograms provide more detailed information about how the data in a dataset are distributed. They are not as good as boxplots for making comparisons, and there's the problem of what interval width to select.

In this section you have looked at histograms and bar charts. You saw that histograms are suitable for use with continuous data, whereas bar charts are used to depict discrete data (which can be either categorical data or numerical data). For bar charts, the convention is to include gaps between the bars, whereas in a histogram the bars touch. With the exception of the illustrative example in Figure 13, all the histograms presented in this module are constructed using equal intervals.

There are other types of statistical charts that are not covered in this module, such as *pie charts*.

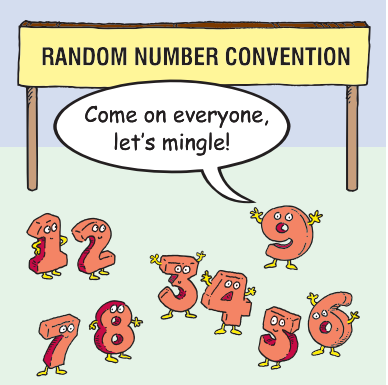
For more information about pie charts, see Maths Help, Module 5, Subsection 2.3.

The next section introduces a new topic, *randomness*, and explores the natural variation that you might expect a sample of random numbers to exhibit.

3 Random numbers

In this section you will use bar charts to explore features in random numbers that arise from chance alone. The aim is to help you to develop a better understanding of features and variations that occur naturally in random data, which will be a useful benchmark against which to compare data arising from a statistical investigation. This is something that you will get an opportunity to do in Section 4, where you will use this approach to test whether or not it can be demonstrated that someone possesses extra-sensory perception (ESP). The key issue underlying all investigations of this nature is to analyse and interpret the observed features in data, and conclude that either:

- yes, there is a real effect here, or
- no, these features could simply be random fluctuations.



3.1 Clusters

When medical case histories are collected together, it is often possible to see a bigger picture about issues such as which categories of patients are becoming ill with what illnesses, and where these illnesses are occurring. In particular, cases of an illness sometimes cluster in certain geographical areas. A problem lies in interpreting such a cluster of ill health, particularly if it occurs close to an unpopular installation such as a phone mast or a nuclear power plant. The key questions here are:

- Is the cluster larger than you might expect?
- If it is, then is its size sufficiently large to lead you to believe that it is more than just a random fluctuation?

Activity 11 *Reading a news story critically*

Have a look at the news story below.

> **Nuclear link to child leukaemia 'cluster'?**
>
> The nuclear industry last night rejected claims of a cluster of children's cancer in North Wales. Radiation expert Chris Busby says the Menai Strait area has a 28-fold rate of child leukaemia compared to the UK average and blames the Sellafield nuclear reprocessing plant in Cumbria in an HTV programme tonight. Dr Busby, of Aberystwyth, who sits on government committees, said: 'There is a 28-fold excess of child leukaemia in Caernarfon over the period 2000–03, three cases, whereas only 0.1 should be expected in comparison with the national average.'
>
> Hywel Trewyn, *Daily Post*, 10 February 2004

(a) What data have been used as evidence in the article?

(b) Can you suggest any problems with the way that the data have been interpreted?

A key question in any analysis of clusters of ill health in particular towns or regions is whether such a cluster is a direct result of some identifiable factor or factors, or whether it is simply a chance occurrence.

In order to make an informed comparison between a cluster of ill health and what you might reasonably expect to happen by chance alone, you need to have an understanding of just what sort of variations you are likely to get due to chance. In this section, therefore, you will be asked to consider just this – what sort of variations tend to crop up naturally through chance alone? Research suggests that many people have a rather poor sense of this, and typically underestimate the extent to which the outcomes of chance events tend to vary and cluster. As a result, they have a tendency to be more surprised and impressed by everyday coincidences and associations than they should be. This subtle but important point was understood nearly two thousand years ago by the Greek essayist Plutarch when he wrote:

> It is no great wonder if, in the long process of time, while Fortune takes her course hither and thither, numerous coincidences should spontaneously occur.
>
> Plutarch (c. AD 46–AD 120)

The main purpose of this section, therefore, is to invite you to explore and develop your own understanding in this area of chance. You will be looking at simple experiments – in statistics these are often referred to as *trials* – and their corresponding outcomes. In general, a **trial** is associated with a number of possible outcomes, but only one of these outcomes can occur at a time. For example, a trial might be the tossing of a coin, while its corresponding outcomes are heads and tails.

You will be asked to use a module software tool to generate some random numbers, and check just how much variation and clustering appears to occur from chance alone. After you have worked through this section, you should be more aware of how variations can arise by chance and be less ready to jump to unjustified conclusions.

3.2 How random numbers vary

Imagine that you have ten identical balls, labelled 0 to 9. They are placed inside a bag and thoroughly mixed around. Now, without looking inside the bag, you pick a ball. Each ball should have an equal chance of being chosen – a condition that ensures that the selection is *at random*. The number of the ball chosen is written down, the ball is replaced, and you repeat this activity a further nineteen times.

This should result in a run of twenty randomly-chosen numbers from 0 to 9. Such lists of numbers are called lists of **random numbers**.

Activity 12 *Making up your own 'random' numbers*

Write down the sort of run of twenty numbers that you think might result from the exercise described above. Don't skip over this, as you are shortly going to be asked to analyse the run of numbers that you think up.

Any made-up run of 'random' numbers might look truly random. But how random is it really?

Here are twenty made-up 'random' numbers:

4 1 9 0 6 3 2 7 8 3 4 6 5 9 3 2 0 8 5 7

And here are twenty computer-generated random numbers (obtained by using the random command on a computer):

9 0 8 3 7 8 8 6 1 1 1 4 6 6 8 1 8 2 9 2

You might be wondering whether the random numbers produced by a computer's random command are truly random. Computer-generated random numbers are often referred to as **pseudo-random**, and there is no guarantee that the computer's random number generator is a perfect model for random selection. However, in practice, they do provide a good match with numbers selected randomly using dice, coins or spinners, or by drawing balls from a bag, in terms of the properties explored in the next few activities. It's not just computers that generate pseudo-random numbers – many calculators have a random number generator.

At first sight, the two runs of numbers above seem rather similar – each is merely a run of digits from 0 to 9 with no special pattern. However, as you will see, the person who made up the twenty numbers actually imposed a greater degree of orderliness than was produced by the computer.

Children in playgrounds use various rhymes to pick people out 'randomly' for games, such as 'One potato, two potato, three potato, four; five potato, six potato, seven potato more'. Of course, these choices aren't really random. If you use the 'One potato...' rhyme, for example, you will always finish on the fourteenth person. (If there are fewer than 14 to choose from, you go round the circle however many times it takes.) But without thinking carefully, it's hard for a child reciting the rhyme to work out in advance who is going to be picked – so the choice is effectively random.

Activity 13 *Checking for number pairs*

Look at the two runs of numbers above and check whether any pairs of consecutive numbers are the same.

Typically, you find that a person making up a run of 'random' numbers tends to avoid having two consecutive numbers the same, whereas the computer may not.

Activity 14 *Checking your own run of numbers for number pairs*

Now check your own made-up run for pairs of consecutive numbers that are the same. Have you tried to avoid repetitions (consciously or subconsciously)?

Here is another check on orderliness.

Activity 15 *Finding the frequencies of the numbers*

For each of the two runs of numbers above, count the frequency of occurrence of each number.

Activity 15 shows that the computer-generated random numbers showed greater variation in the frequency counts than the made-up numbers.

Activity 16 *Finding the frequencies in your own run of numbers*

Now check the frequencies of the numbers in your own made-up run. Did each of your numbers come out with roughly the same frequency, or did the frequencies fluctuate widely, as for the computer-generated run of numbers?

When people make up their own 'random' numbers, they tend to produce numbers with much narrower variations in frequency than is found in random numbers generated by a computer or calculator. This contrast between made-up and computer-generated random numbers is even more evident when the frequencies are displayed in bar charts. Figures 14 and 15 show the bar charts for the two runs of numbers on page 29. The numbers 0 to 9 are shown along the horizontal axis, and the heights of the bars indicate the frequencies with which the numbers occur.

Note the gaps between the bars, which are a feature of bar charts.

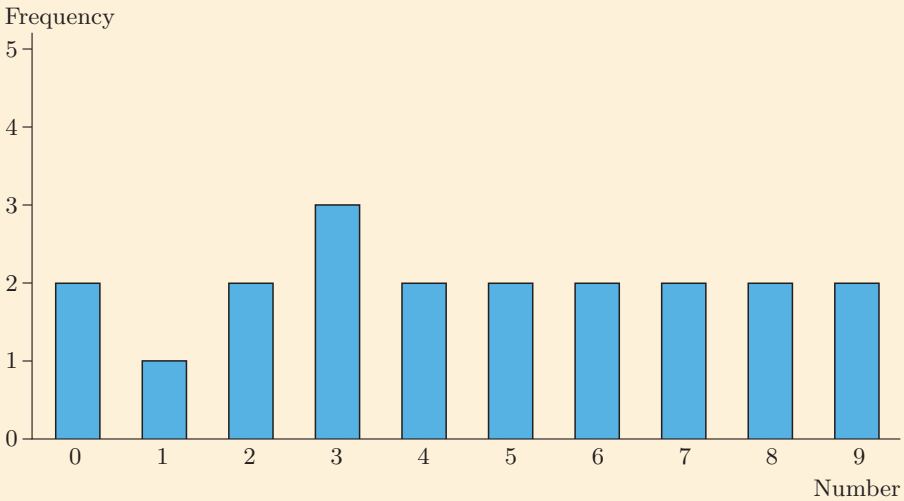


Figure 14 A bar chart showing the frequencies for the made-up run

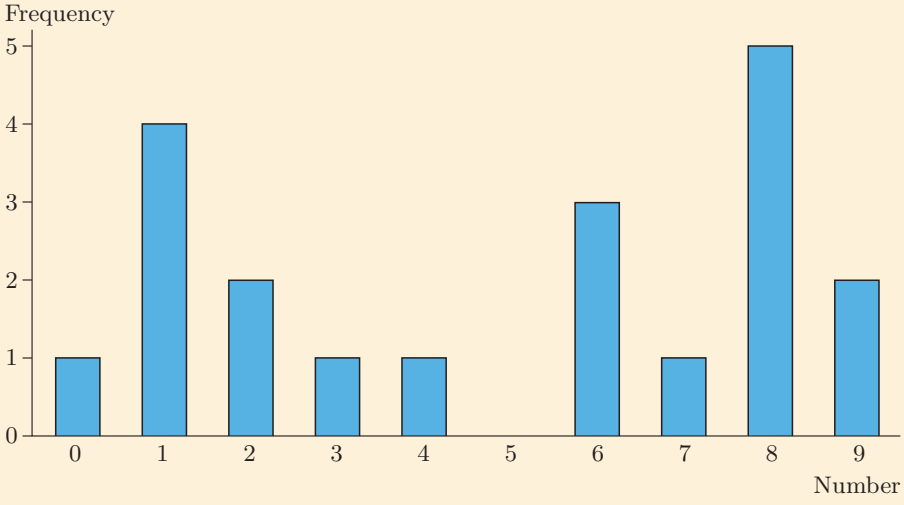


Figure 15 A bar chart showing the frequencies for the computer run

The point of the activities that you have done so far in this subsection was to indicate that random events may produce outcomes that fluctuate more widely than most people expect. However, it is difficult to show this convincingly on the basis of a single run of twenty computer-generated numbers. You really need to generate more than one run of random numbers from your computer if you are to develop a good sense of the disorderliness of random events.

Note that one important characteristic of a run of random numbers is that the selection of each new random number does not depend in any way on which numbers were selected previously in the run. You can see that this is true when the numbers are generated by choosing balls numbered 0 to 9 from a bag, in the way described earlier. We say that each new selection is **independent** of the previous selections.

At this point, it is useful to distinguish between uniform and non-uniform random numbers.

The random numbers that you have been thinking about in this subsection are *equally-likely* random numbers. For example, if you generate the random numbers by choosing balls from a bag in the way described, then you are just as likely to choose any of the balls as any of the others, so the random numbers generated are all equally likely. Equally-likely random numbers are also known as **uniform random numbers**.

The dice used in casinos are known as *precision dice*. They are carefully manufactured to ensure that the different numbers come up equally often. To ensure that they are perfectly balanced, the spots are completely filled with material of the same density as the rest of the cube, and often the plastic material is translucent, so that weights cannot be hidden inside!

Now imagine that there are eleven balls in the bag, the same ten as before, together with an extra ball labelled 0, making two balls labelled 0 altogether. Now when you generate a run of random numbers by choosing balls out of the bag, you are twice as likely to generate a 0 as you are to generate any of the other numbers. So the random numbers produced in this way are not equally likely; that is, they are *non-uniform*. Similarly, rolling a die whose faces are labelled 1, 1, 2, 2, 3, 4, for example, will produce non-uniform random numbers.

All of the random numbers generated in this unit are uniform random numbers.

Investigating variation in random numbers

In the rest of this subsection you will have the opportunity to use a module computer resource called ‘At random’ to generate some runs of random numbers for yourself. The process of generating a run of random numbers in this way will be referred to as a **simulation**, since it simulates what might happen if, for example, you were to pick numbered balls out of a bag in the way described earlier. Simulations involving computer-generated random numbers are often used in statistics, to show what might happen in a real situation that involves randomness.

In the next activity, you are asked to use the ‘At random’ software to generate sixty random numbers between 0 and 9. Remember that even though the numbers are equally likely to occur, they will almost certainly not crop up equally often in practice. The objective of this activity is for you to get a sense of how much natural variation there is, particularly with relatively small runs of numbers such as the runs of sixty numbers investigated here, and to describe this variation.

At random

Activity 17 *Investigating variation in a run of random numbers*

Open the 'At random' software.

(a) Check that the settings at the bottom of the right-hand panel are as follows.

Generate values between: 0 and 9

Number of values (per run): 60

Number of runs: 1

Then click 'Go' to run the simulation. The software generates 60 random numbers and draws a bar chart showing the frequencies.

(b) Write a few sentences describing the variation in the frequencies of the numbers.

As you saw in the comment on Activity 17, for the particular simulation shown there, the most frequent number occurred six times as often as the least frequent number. You might be wondering just how typical this degree of variation is. For example, if you calculate the ratio of maximum frequency to minimum frequency, based on your own results in Activity 17, is it as large as 6?

In the next activity, you are asked to run the simulation several more times to investigate the variation in the frequencies of the numbers. In particular, you are asked to look at the ratio of maximum frequency to minimum frequency in each run, which is calculated automatically by the 'At random' software.

At random

Activity 18 *Investigating variation in more runs of random numbers*

(a) Look again at the results of the simulation that you carried out in Activity 17. (If you no longer have the results, then carry out another simulation, with the same settings.)

In the table in the left-hand panel, in the column headed 'Max/Min', you will find the ratio of maximum frequency to minimum frequency. What is its value for your simulation?

If the minimum frequency is 0, then the ratio cannot be calculated and a dash is displayed in the table. If this happens, then just continue with part (b), or carry out another simulation, if you wish.

(b) Now set 'Number of runs' to 10 (keeping the other settings the same as in Activity 17), and click 'Go'. The software generates 10 runs of random numbers. The maximum frequency, minimum frequency and ratio of maximum frequency to minimum frequency for each of the 10 runs are displayed in the left-hand panel. The bar chart displays the results of all 10 runs put together.

Write a few sentences describing your findings, and indicate what they suggest in terms of gaining an insight into the variability of random numbers.

The main point to emerge from your investigation in this subsection can be summarised as follows.

Suppose that you have a fairly small run of random numbers, such as a run of sixty random numbers. Then even if all the possible numbers are equally

likely to occur, the degree of variation in frequency between the numbers is likely to be quite large, and probably larger than most people expect.

However, what happens if you look at larger runs of random numbers? This is the topic of the next subsection.

3.3 Larger runs of random numbers

So far, you have been looking at fairly small runs of random numbers (with no more than 60 numbers). What happens when the number of random numbers in a run is greatly increased? For example, would there be as much disorderliness if the number of random numbers were increased to, say, 300 or 10 000 or even 100 000? Fortunately, these are questions that are easily investigated using the 'At random' software.

At random

Activity 19 *Increasing the number of random numbers*

Return to the 'At random' software.

(a) Choose the following settings.

Generate values between: 0 and 9

Number of values (per run): 300

Number of runs: 1

Run the simulation several times (by clicking 'Go' each time) and compare the overall shape of each bar chart, and each ratio of maximum frequency to minimum frequency, with the typical shapes and ratios that you got when the number of values was 60.

(b) Repeat part (a) with the number of values increased to 10 000. Compare the shape of each bar chart, and each ratio of maximum frequency to minimum frequency, with what you saw previously.

(c) The maximum number of values per run that the software will accept is 10 000, but you can see the shape of the bar chart that results from a run of 100 000 values by leaving the number of values per run at 10 000 and increasing the number of runs to 10. The corresponding ratio of maximum frequency to minimum frequency is then given in the bottom row of the table.

Make this change to the settings, then run the simulation several times and compare the shape of each bar chart, and each ratio of maximum frequency to minimum frequency, with what you saw for fewer random numbers.

Activity 19 illustrates some important facts about the nature of random numbers in which all the possible numbers are equally likely to occur – that is, uniform random numbers. These facts are set out in the box below.

Uniform random numbers

- When a fairly small run of uniform random numbers is chosen, the degree of disorderliness in the numbers is often surprisingly high.
- With larger runs, the frequencies tend to settle down and become approximately equal.

3.4 Using random variation as a basis of comparison

This final subsection in this section explains why it is useful to have a good grasp of randomness and the effects that it can produce. For example, some medical conditions such as leukaemia or high blood pressure sometimes appear to occur in identifiable clusters in particular regions of the country. In recent years, there have been many newspaper scare stories about possible causes (mobile phone masts, nuclear power plants, electricity lines, and so on). But a crucial question here is whether such a cluster is caused by some external factor or whether it is simply one of the high frequencies that occur naturally with random events.

At random

Activity 20 *Cause or coincidence?*

Imagine that in a particular year there is a total of 60 known cases of a particular type of leukaemia, spread randomly over six similar towns, all with roughly the same population.

(a) If there are no differences in the circumstances of the towns, how many cases of the disease, *on average*, might each town expect to have in that year?

(b) Think of each town as labelled with a different number from 1 to 6. Use the 'At random' software to generate sixty random numbers between 1 and 6 inclusive, and note how many times each of the numbers occurs. Explain how the random numbers that you have generated simulate the possible distribution of the leukaemia cases.

(c) Suppose that in the actual distribution of cases, one of the towns is unlucky enough to account for 17 of the 60 cases, and this town happens to be closest to a nuclear installation.

Set the 'At random' software to carry out the simulation ten times (by setting the number of runs to 10), and look at the results. On the basis of these simulations, do you think that the citizens from this town have a cause for concern?

Running the simulations in Activity 20 will produce different results every time. A member of the module team ran the ten simulations (see Figure 16) and found that in one of them, one of the numbers had a frequency of 19.

Figure 16 Ten simulations of the leukaemia scenario in Activity 20

So if you consider the frequencies of leukaemia cases in the six towns, then even a frequency as high as 19, which is well above the average value of 10, could arise by chance.

So is there a cause for concern for the town with a frequency of 17 cases of the disease? You might conclude that the cluster of cases experienced by this town is small enough to be just a chance occurrence. Although this may be the explanation, it would be sensible for the town authorities to monitor the number of cases carefully over subsequent years to check whether it continues to be higher than expected, in which case there would be a much greater cause for concern. They might also, particularly if the frequency continues to be high, seek scientific advice about whether the cluster of cases might be causally linked to the nuclear installation, or perhaps to some other factor.

So, in general, a cluster of ill health occurring in a particular region *may* imply a possible cause-and-effect explanation based on factors peculiar to that region. However, such a conclusion should be viewed with caution: it must be weighed against a possible alternative explanation that the cluster is just an extreme result due to random fluctuations.

In this section you have used the random number generating software 'At random' to explore variation in random numbers. The point of this exercise was to get a better sense of just how widely fluctuating this variation can be.

In statistics, knowing the extent of random fluctuations provides a useful benchmark against which to interpret experimental data – an idea that you will revisit in Section 4, where you will be asked to carry out an investigation on testing for extra-sensory perception (ESP).

4 Case study: a statistical investigation of the paranormal

In this final section, you are asked to take some of the statistical skills that you have learned in Unit 4 and in the first three sections of this unit, and apply them to an investigation that involves making a statistical judgement. The context is about how you might decide whether or not *extra-sensory perception* (ESP) exists.

Extra-sensory perception, if it exists, is the ability to acquire information by paranormal means – in other words, by a means that does not depend on any known physical sense or any deduction from previous experience.

As part of this investigation, you will be asked to carry out a short experiment to gather data. Bear in mind that the point of the case study is for you to gain insights about the statistical method of investigation. You can best do this by engaging with the experiment rather than simply reading about the statistical investigation in the abstract. Also, you will find that the experience of having collected, analysed and interpreted *your own data* will be helpful to refer to when you are working through some of the more complex ideas in the section.

You may possibly have an opinion about ESP already – you may strongly disbelieve in its existence, you may strongly believe in it, or you may be undecided. However, it is important to understand that the point of this section is not to prove or disprove the existence of ESP, but rather to demonstrate the general process of statistical decision making for questions where absolute proof is difficult to achieve.

A key role of the statistician when investigating some phenomenon is to abandon all personal opinions or prejudices and look only at the evidence. If a statistician is asked to investigate, say, whether a particular drug is effective or whether a certain factory poses health risks or whether ESP exists, then the methodology is always the same. The investigation should begin with the assumption that the phenomenon does not exist (that is, the drug isn't effective, the factory poses no health risks, ESP doesn't exist). This is the starting assumption of 'no difference', sometimes referred to as the **null hypothesis**. Bearing in mind the possibility that occasional fluke results will always happen randomly, the statistician must then examine the data and ask the question: How much evidence of difference would I need in order for me to abandon my null hypothesis and conclude that the phenomenon *does* exist?

You are asked to use this methodology to explore the question of the existence of ESP in a systematic manner, using the four-stage PCAI investigation cycle described in Unit 4, and summarised in the box below.

The four stages of a statistical investigation

Stage 1 **P**ose a question
Stage 2 **C**ollect relevant data
Stage 3 **A**nalyse the data
Stage 4 **I**nterpret the results

4.1 Stage P: posing a question about ESP

The first stage in any statistical investigation is to clarify the question of interest. Often this involves maintaining a compromise between what you really want to know and what it is possible to find out with the statistical tools at hand. Here are some general issues that are an important backdrop to any investigation.

Starting an investigation

- Think carefully about the wording of the question of interest. Make sure that it is specific and unambiguous.
- Check that the question can be answered by undertaking a doable investigation. This requires anticipating the possible stages that will lie ahead – what data will you need and how will you collect them (the 'C' stage); how will you process the data (the 'A' stage); and is it likely that your choice of analytical tools will give you an answer to your question (the 'I' stage)?

In this section you are asked to work with the following question:

> Does a person whom I know have extra-sensory perception (ESP)?

This is a very open-ended question, and a great deal of work would be needed to investigate it thoroughly, so in this section you are asked to narrow it down and think about evidence of ESP in a particular context – namely, how good a person is at guessing numbers on cards that are hidden from him or her. A more specific version of the question above, to suit this context, is:

> Does a person whom I know show evidence of ESP when guessing hidden numbers on cards?

You will be asked to invite two friends or family members to be your *subjects*, and to test their card-guessing abilities using cards something like those shown in Figure 17. Playing cards would be suitable – for example, the Ace, 2, 3 and 4 of Spades.

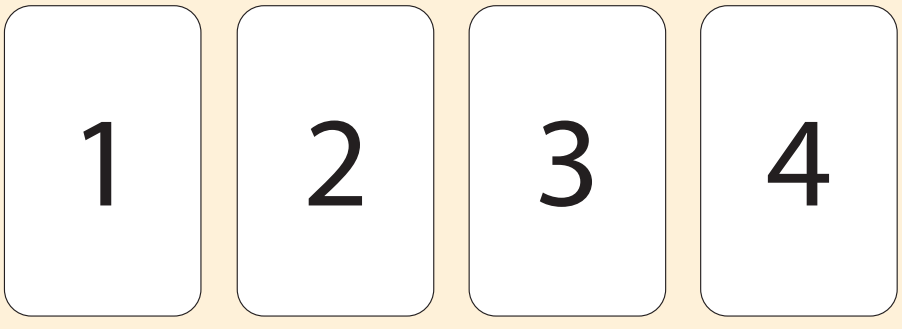


Figure 17 Cards numbered 1 to 4

As part of the investigation, you will explore what 'evidence of ESP' might mean, by considering computer simulations of results that might be obtained just by chance.

4.2 Stage C: collecting the card-guessing data

At the 'C' stage of any investigation, you need to collect relevant data. In the investigation here, an experiment must be designed to gather data to answer the question posed in the last subsection.

Here is the experiment that you will be asked to carry out later in this subsection.

Instructions for the card-guessing experiment

1. Identify two people who are willing to be tested to see how good they are at guessing numbers on cards. They can be reassured that you will require only five minutes of their time.
2. Prepare four cards, numbered 1, 2, 3 and 4, on one side, as shown in Figure 17 (or use playing cards).
3. Prepare a data collection sheet. You can either print out the sheet in the Unit 11 resources section of the module website, or draw up your own version. For each subject, you will need a table similar to the one shown (completed) in Figure 18, with three columns, headed 'Card', 'Guess' and 'Actual', and twenty rows. You will also need a 'Score' box for each subject, to record the number of correct guesses out of 20.
4. Test your first subject as follows. Shuffle the four cards thoroughly, choose one at random and, without looking at it yourself, lay it face down. The subject now says what they think the number is; write this in the first row of the column headed 'Guess'. Then look at the card and write down its number in the first row of the column headed 'Actual', without the subject seeing what you are writing. Put the chosen card back to rejoin the other three cards.

   Repeat this for twenty chosen cards, and then write the number of correct guesses in the Score box. The completed table and Score box should look something like those in Figure 18.
5. Repeat step 4 for your other subject.

Subject A

Card	Guess	Actual	
1	2	1	
2	4	4	✓
3	1	2	
4	3	4	
5	2	3	
6	1	1	✓
7	4	2	
8	3	4	
9	4	4	✓
10	2	4	
11	3	4	
12	3	4	
13	2	1	
14	1	1	✓
15	1	3	
16	3	3	✓
17	4	4	✓
18	2	1	
19	1	2	
20	3	1	

Score 6

Figure 18 An example of a completed table and Score box for the card-guessing experiment

Once you have completed the experiment, you will need to try to make a judgement from your subjects' scores as to whether they have shown evidence of possessing ESP.

An important issue in experimental design is how to ensure that a test is 'fair' – the experiment should test what you intend to test! A useful approach to thinking about this is to try to anticipate how a cheat might attempt to subvert any test that you devise, and then create 'laboratory conditions' to counteract such subversion. Activity 21 asks you to think about how the test in step 4 of the experiment above could be refined to make sure that it is fair.

Activity 21 *Checking whether the test is fair*

(a) What would count as cheating in the test described?

(b) How might you make it difficult for a subject to cheat?

In addition to minimising the possibility of cheating, your test must of course be open to the possibility that your subject has ESP. With some ESP experiments, it is possible that the circumstances of the actual test may affect the level of performance. For example, some people find tests stressful and they do not perform as well as they might have done. So it is important to provide a relaxed environment. Try to set aside a reasonable amount of time, and if possible ensure that you and your subjects will not be disturbed.

A common complaint of 'psychics' who fail to perform well when tested is that they have underachieved because of the 'sterile' atmosphere of the 'laboratory conditions'.

The next activity asks you to conduct the card-guessing experiment. If you find it impossible to do this (perhaps because you are unable to find suitable subjects who are willing and able to help), then you can instead carry out a different version of the experiment, which is available on the module computer resource 'By chance alone'. In this version, the numbers on the cards are computer generated, and *you*, rather than two different subjects, are tested on how good you are at guessing the numbers. The instructions for this version of the experiment are given in the activity.

By chance alone

Activity 22 *Conducting the experiment*

Either

Carry out the experiment with your two subjects as described on the opposite page. Keep a note of the two subjects' scores out of twenty, as you will need them in the rest of the section.

or

Carry out the 'self-test' version of the experiment, as follows.

(a) Open the 'By chance alone' software, make sure that the 'Card experiment' tab is selected, and click 'Self test'.

(b) The backs of twenty cards are shown on the left of the software display. Guess the number on the first card (1, 2, 3 or 4). Then, in the table headed 'Subject A', type your guess in the first cell of the 'Guess' column, and press 'Enter'. The first card turns over and its number is displayed in the first cell of the 'Actual' column. The cursor moves down, ready for you to guess the number on the second card.

(c) Enter your guesses for all twenty cards, and click 'OK' in the dialogue box that appears. The software displays the number of correct guesses in the Score box at the bottom of the table.

(d) Now the backs of a further twenty cards are shown. Enter your guesses for the numbers on these cards in the table headed 'Subject B', and click 'OK' in the dialogue box that appears.

(e) Make a note of the two scores out of 20 that are displayed in the Score boxes at the bottom of the tables, as you will need them in the rest of the section. If you plan to continue your study session, then leave the 'By chance alone' software open and unchanged, as you will be able to continue with it in the next activity. If you do not plan to continue, then just keep your note of the two scores.

(There are no comments on this activity.)

Whether you carried out the experiment on two subjects, or the alternative version in which you tested yourself, you should now have two scores out of 20, one for Subject A and the other for Subject B. You are now going to analyse these results to see just how well your subjects performed, by comparing their two scores with the scores that might be achieved by random guessing.

4.3 Stage A: analysing the card-guessing data

With the two scores for Subject A and Subject B in mind, you now need to think about how successful your two subjects really were. For example, suppose that one of them scored 7 out of 20. Is that a lot or a little, and by what criterion can you make that judgement?

As you saw in Section 3, the most useful way to judge results like these is by comparing them to the answer to the question: What sort of results would you expect to get by chance alone? Clearly if you simply guess the answers, then you are likely to get *some* right by chance. Evidence of possessing ESP would require a 'better-than-chance' performance.

As there are four cards in the test, a subject who guesses randomly has a one-in-four chance of guessing any particular card correctly. So he or she should get about 5 out of 20 guesses right, on average. But if you were to look at the scores out of 20 for a large group of subjects guessing randomly, would these scores tend to be bunched between, say, 4 and 6, or is the variation likely to be much wider – say between 1 and 9? It is important for you to know this if you are to identify a 'better than chance' performance by one of your subjects. To answer this question, some investigation is needed into scores that can be obtained by guessing randomly. This can be done by using simulations.

In this subsection you will use the module software 'By chance alone' to carry out suitable simulations. In each simulation, the software generates two lists of twenty random numbers from 1 to 4. One list of random numbers simulates the random guesses that are made, and the other list simulates the actual numbers on the cards. The software counts the number of 'correct guesses' to give a score out of 20. It can run the simulation many times, to give an indication of the variation in the scores.

In the next activity you will look at what happens if the simulation is run 100 times.

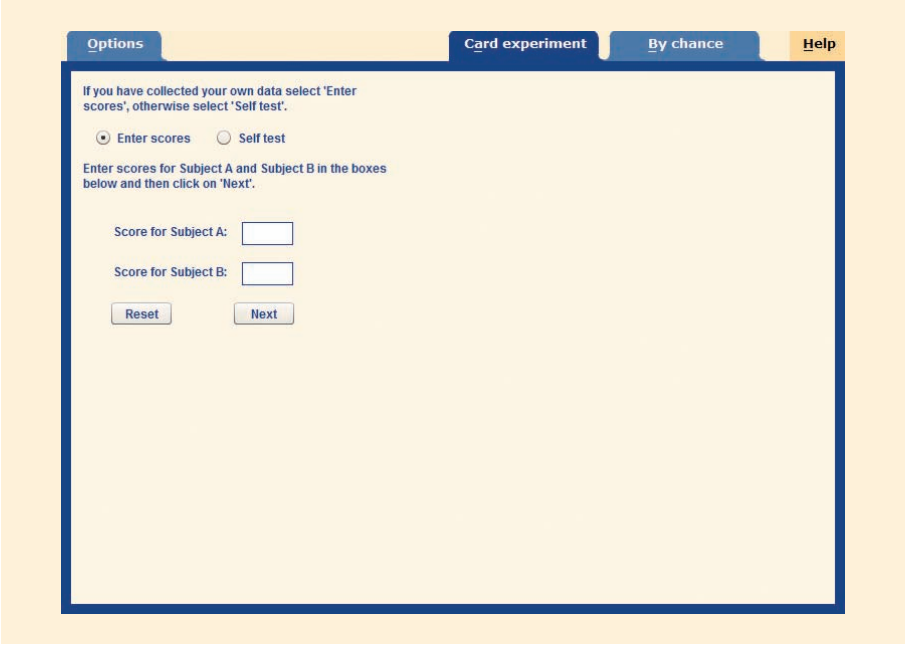
By chance alone

Activity 23 *Simulating scores obtained by guessing*

Open the computer resource 'By chance alone', if it is not already open.

(a) If you have left this software open and unchanged since you carried out the previous activity, then click the 'Next' button at the bottom right of the software display, and go on to part (b).

Otherwise, make sure that the 'Card experiment' tab is selected, and click 'Enter scores'. You should then see the screen shown below.



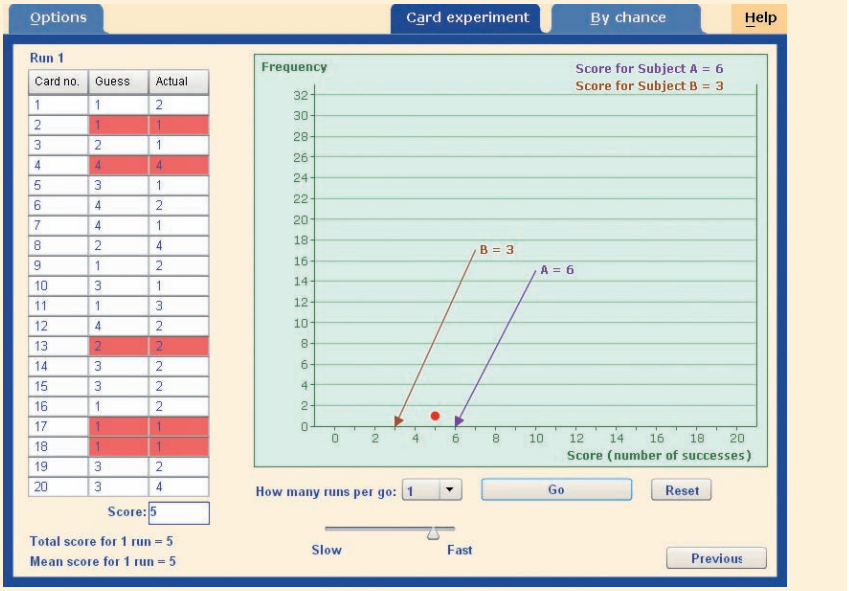
Enter the scores for Subjects A and B that you obtained in the previous activity. Then click the 'Next' button.

(b) You should now see a screen similar to the one below, but with the scores of your two subjects displayed at the top right and marked on the chart.

(c) Check that 'How many runs per go' (below the chart) is set to 1. Then click the 'Go' button. Several things will happen.

- Two lists of twenty random numbers are generated, one in the column headed 'Guess', to represent the twenty random guesses, and the other in the column headed 'Actual', to represent the actual numbers on the cards. All the 'correct guesses' are highlighted in red, and the total number of correct guesses is displayed in the Score box below the table.
- Once all twenty rows have been filled, a red dot appears on the chart, to represent the score. For example, if the score is 5, then

the red dot appears above the point on the horizontal axis that represents 5, as shown in the screenshot below.



Run the simulation (by clicking 'Go') several times, in order to understand what is going on.

(d) Now change the number of runs per go to 10, click 'Go', and watch what happens. The simulation is run 10 times. The 'Run' number at the top left of the software display tells you which of the 10 simulations is currently being run. After each run, the score is plotted on the chart, so a dotplot of the scores is built up. The most recent score appears as a red dot.

The dotplot displayed by the 'By chance alone' software has a vertical scale marked, so that the frequencies are easier to read off.

(e) Now run 50 simulations (by increasing the number of runs per go to 50 and clicking 'Go'), and finally run 100 simulations.

To speed things up, use the Slow–Fast slider near the bottom of the software display.

In the activity above you used the 'By chance alone' software to generate 100 scores typical of those that might be obtained simply by guessing the numbers on the cards. The dotplot of these scores gives you a sense not only of where the scores are centred – their location – but also of how they are spread.

For example, Figure 19 shows a dotplot of 100 scores obtained from the software. In this dotplot, the scores seem to be centred roughly at 5, as you would expect, but there are scores as low as 0 and as high as 11.

The set of scores in Figure 19 provides a useful, if rudimentary, benchmark against which to test your subjects' scores. For example, if one of your subjects scored 7 out of 20, say, then you would probably not consider this to be possible evidence of ESP, since this score seems to be well within the range of scores that are obtained simply by guessing. On the other hand, if your subject scored 9, say, then you might consider this to be of more significance, because, as you can see from the dotplot, only four of the 100 scores were as high as 9 or more (there were three scores of 9, no scores of 10, one score of 11 and no scores any higher than this).

Figure 19 A screenshot of the 'By chance alone' software, showing typical scores for 100 runs of the card-guessing test if the subject simply guesses the numbers on the cards

So you now need to consider the question of how high a score has to be if you are to regard it as showing evidence of ESP. You would want the score to be high enough to ensure that there is only a very small chance of obtaining a score that high or higher simply by guessing the numbers on the cards.

Let's suppose that we want to regard a score as showing evidence of ESP if there is only a 5% chance, or less, of obtaining a score that high or higher by guessing. In other words, we will regard a score as showing evidence of ESP if it falls into the top 5% of scores that might be obtained by guessing. You might think that this percentage seems too high, but let's stick with it for the moment, to illustrate the ideas.

There are 100 scores in Figure 19, so the top 5% of scores are the top five scores. You cannot identify 'the top five scores' exactly, because they would have to be the score of 11, the three scores of 9 and one of the scores of 8 – but you can't say that one of the scores of 8 is higher than any of the other scores of 8.

However, you can say that a score of 9 or more is definitely in the top 5% of scores. So, if you are regarding a score as showing evidence of ESP if it falls into the top 5% of scores, and you are judging by Figure 19, then you would regard a score of 9 or more as showing evidence of ESP. You would not regard a score of 8 as showing evidence of ESP, because you cannot say that it falls into the top 5% of scores. All that you can say about a score of 8 is that it falls into the top 9 of the 100 scores – that is, into the top 9% of scores. Similarly, you would not regard any score smaller than 8 as showing evidence of ESP.

The top nine scores are the five scores of 8, the three scores of 9 and the score of 11.

If you think that a higher level of evidence is needed than the '5% criterion' described above, then you might instead use a '1% criterion': you would regard a score as showing evidence of ESP only if it falls into the top 1% of scores obtained by guessing. In this case, if you are judging by the dotplot in Figure 19, you would regard only scores of 11 or more as

showing evidence of ESP. You might want to set a level of evidence even higher than this – we'll come back to the question of how high the level should be set later in the section.

A weakness of the analysis above is that the dotplot in Figure 19 is based on only 100 scores typical of those that might be obtained by guessing – really too few to provide a very robust test criterion. In the next activity, you are asked to use the 'By chance' tab of the 'By chance alone' software to generate many thousands of scores, which will provide a much better benchmark against which to judge your subjects' scores. The software displays the results as a bar chart rather than a dotplot.

By chance alone

Activity 24 *Running many simulations*

Open the 'By chance alone' computer resource if it is not already open, and select the 'By chance' tab.

(a) Check that the three settings at the top left of the software display are set as follows:

Success probability: 0.25

Sample size n: 20

Number of runs: 100

The 'success probability' of 0.25 corresponds to the fact that one quarter of random guesses of a number on a card are expected to be correct, on average. This is because there are four possibilities for the numbers on the cards.

The 'sample size' of 20 corresponds to the fact that each simulation mimics a test in which the numbers on 20 cards are randomly guessed, so the scores obtained are scores out of 20.

The 'number of runs' refers, as before, to the number of times that the simulation is run. So, for example, setting 'number of runs' to 100 allows you to generate 100 scores typical of the scores that might be obtained by guessing.

(b) Check also that the box at the bottom left of the software display, labelled 'Cumulate data', is *not* ticked.

Now click 'Go'. The software generates 100 scores and displays the results as a bar chart.

Click 'Go' a few more times to get a feel for what is going on. What do you notice about the shapes of the bar charts?

(c) Set the number of runs to 100 000, and click 'Go'. The software generates 100 000 scores and displays the results as a bar chart.

Now click 'Go' a few more times. What do you notice about the shapes of the bar charts this time?

The very large number of scores that you can obtain from the 'By chance' tab of the 'By chance alone' software gives you a much better benchmark against which to judge the success, or otherwise, of your two subjects in the card-guessing test. Remember that the scores generated by the software simulate the kind of scores that you would expect to be obtained just by guessing. In the next activity you are asked to use a large number of scores to work out the minimum score that you would regard as showing evidence of ESP if you are using the '5% criterion', and similarly if you are using the '1% criterion'.

Activity 25 *Finding which scores satisfy the criteria*

By chance alone

Use the 'By chance alone' software, with the 'By chance' tab selected.

(a) Check that the three settings at the top left of the software display are set as follows:

Success probability: 0.25

Sample size n: 20

Number of runs: 100 000

Also tick the box labelled 'Cumulate data' at the bottom left of the software display. The effect of this is that, each time you click 'Go', the new scores generated are put together with all the previous scores, to give you a larger and larger number of scores. The total number of scores represented in the bar chart is displayed on the left of the software display, labelled as 'Total number of runs'.

Click 'Go' repeatedly until the shape of the bar chart no longer seems to change.

(b) Now click 'Shade score' at the left of the software display, and change the number in the box after the words 'Shade score' from 5 to 8.

The effect of this is that the bars corresponding to scores of 8 or more are shaded, and the percentage of scores that are 8 or more is displayed above the 'Go' button, to three significant figures.

Click 'Go' a few more times – this will generate some more scores – to check whether the percentage of scores that are 8 or more seems to have settled down to a fairly consistent value. You should find that it settles down to 10.2% (to 1 d.p.).

(c) Use the 'Shade score' feature to complete the table below. Each time you change the number in the 'Shade score' box, click 'Go' a few more times to check that the percentage of scores seems to have settled down to a fairly consistent value. Round the percentages to one decimal place.

Score	Percentage of scores that are this high or higher (to 1 d.p.)
8	10.2%
9	
10	
11	
12	

(d) Suppose that you want to regard a score obtained by a subject in the card-guessing test as showing evidence of ESP only if it falls into the top 5% of scores that would be obtained by guessing. Use your answers to part (c) to determine the minimum score that would count as showing evidence of ESP.

(e) Suppose now that you want to regard a score obtained by a subject as showing evidence of ESP only if it falls into the top 1% of scores that would be obtained by guessing. Use your answers to part (c) to determine the minimum score that would count as showing evidence of ESP in this case.

The minimum score that falls into the top 5% of scores obtained by guessing is referred to as the 5% **critical value**, and any score greater than or equal to this value is said to lie in the 5% **critical region**. Similarly, the minimum score that falls into the top 1% of scores is referred to as the 1% critical value, and any score greater than or equal to this value is said to lie in the 1% critical region, and so on.

Before you can make a judgement about how your two subjects performed in the card-guessing test, you need to decide what level of evidence might be appropriate for regarding a score as showing evidence of ESP. Is either the 5% criterion or the 1% criterion good enough, for example?

The next activity should help you to think about this question.

Activity 26 *How many subjects might achieve a score in the critical region?*

(a) (i) By using your answers to Activity 25, write down the 5% critical value for the card-guessing test. Hence, by looking at your answer to Activity 25(c), write down the percentage of subjects taking the card-guessing test that you would expect to achieve a score in the 5% critical region *purely by chance*.

(ii) Now consider all the subjects taking the card-guessing test in a given presentation of MU123. Work out an estimate for the number of these subjects that you would expect to achieve a score in the 5% critical region purely by chance. You can make the following (very approximate) assumptions.

- The number of students in a single presentation of the module is about 2000.
- The proportion of students who are able to find two subjects who agree to participate is about 90%.

(b) Repeat part (a) for the 1% critical value and region rather than the 5% critical value and region.

(c) Do you think that either the 5% criterion or the 1% criterion is strict enough for you to decide whether someone has shown evidence of ESP?

From what you saw in Activity 26, you would probably not consider either the 5% criterion or the 1% criterion to be a strict enough criterion for you to decide whether someone has shown evidence of ESP.

Instead, you would probably want to choose a criterion with a much smaller percentage. There is no correct answer for what the appropriate criterion should be: it is for the investigator to decide what seems appropriate.

For example, consider the '0.0001% criterion'. With this criterion, you would regard a score as showing evidence of ESP if it lies in the 0.0001% critical region. This means that the chance of obtaining a score that high or higher just by guessing is less than 0.0001%. If, say, 3600 subjects are tested in a given presentation of MU123, as estimated in Activity 26, then the number of these subjects that you would expect to obtain a score in the 0.0001% critical region purely by chance is less than

$$0.0001\% \text{ of } 3600 = 3600 \times \frac{0.0001}{100} = 0.0036.$$

So it seems very unlikely that any subject at all will achieve a score in the 0.0001% critical region purely by chance, and so you might be willing to accept a score in this region as evidence of ESP. As you can check by using

the 'By chance alone' software if you wish, a score has to be 16 or more to lie in the 0.0001% critical region.

Activity 27 *Analysing your subjects' results*

Do either of your subjects' scores lie in the 0.0001% critical region? That is, are either of their scores greater than or equal to 16?

(There are no comments on this activity.)

The 'By chance alone' software uses a type of scientific notation to display very small numbers. For example, it would display the percentage

0.0000387%

as

3.87e-5%,

which means

$(3.87 \times 10^{-5})\%$.

The method that you have seen in this subsection for analysing subjects' scores – checking whether they are in the top so-many percent of results that might be obtained just by chance – is an important method in statistics. For example, the actual results in an investigation, such as a drug trial, might be regarded as showing evidence of a real effect if they are in the top 5% or top 1% (or some other percentage) of results that might be obtained just by chance. Much statistical decision making is based on methods of this kind.

4.4 Stage I: interpreting the ESP results

The final stage in the PCAI statistical investigation cycle is to interpret the results that have been obtained, in order to answer the question posed. In this investigation, the question was:

> Does a person whom I know show evidence of ESP when guessing hidden numbers on cards?

Towards the end of the last subsection you were asked to look at whether either of your subjects' results lies in the 0.0001% critical region – you saw that a score lies in this region if it is 16 or more. You may or may not have found that at least one of your subjects' scores lies in this region. But how should your results be interpreted?

If neither of your subjects achieved a score in the 0.0001% critical region – that is, if both of them scored 15 or less – then you might interpret their scores as 'insufficient to show evidence of ESP'. However, if a subject scored more than about 10, say, then you might point out that his or her score is quite high in comparison to the typical scores that are obtained by chance, and further investigation might be warranted.

What if one of your subjects achieved a score in the 0.0001% critical region – that is, what if he or she scored 16 or more? Although the chance of obtaining a score this high just by guessing is less than 0.0001%, it is still possible, of course, that a score this high *could* be obtained just by guessing. So, if one of your subjects achieved a score of 16 or more, then you might want to interpret their score as possible evidence of ESP, with the proviso that this conclusion is not a certainty.

In general, unlike the situation for much of pure mathematics, statistics offers no certainties. For this reason, decisions made on the basis of interpreting data sometimes turn out to be wrong.

Even an unlikely event can occur many times if you give it enough chances to happen. You might think that tossing 20 heads in a row with a coin would be an extremely unlikely event – and it is: the odds are about one in a million. But if all 60 million people in the UK tossed a coin 20 times, then you would expect about 60 of them (60 million divided by 1 million) to toss 20 heads in a row. Individually, these people are likely to feel that something spooky has happened, but if we look at the population as a whole, then we see that their results are no more than we would expect by chance.

Activity 28 *What if one of your subjects appeared to demonstrate ESP?*

Suppose that one of your subjects achieved a score of 16 or more on the card-guessing test. How might you proceed with regard to him or her?

Of course there are many ways to work with the card-guessing test, and to judge the results, other than those suggested in this section. For example, rather than looking for a score of 16 or more, say, on one test, you might ask each subject to take the test several times, and you might regard their scores as showing evidence of ESP if *each* of the scores is fairly high, though not necessarily as high as 16. You could decide, perhaps by using simulations, what you would want the minimum score on each test to be in order to ensure that the chance of achieving a score that high in each of several tests simply by guessing is very small.

The module software 'By chance alone' is not set up to carry out simulations for the approach described here.

Parapsychology: science or pseudo-science?

Parapsychology remains a hugely controversial subject. Some parapsychology experiments have appeared to demonstrate evidence of ESP, but often these results can be explained by flaws in the experimental design or its execution, with fraud and collusion also sometimes coming into the frame.

An interesting player in the world of parapsychology is James Randi (Figure 20). Styling himself 'magician, sceptic and writer', Randi has spent decades challenging those who claim to possess paranormal powers to demonstrate objective proof of their abilities under scientific testing criteria. To date, the prize money offered for such proof by the James Randi Educational Foundation, which currently stands at $1 000 000, has not been won. In fact, no claimant has yet progressed past the preliminary test, which is set up on terms agreed in advance by both Randi and each claimant.

Critics of Randi argue that he has set himself up as judge and jury, and that a true prize would be controlled by an independent panel of neutral judges that would determine whether or not an applicant had truly demonstrated psychic powers.

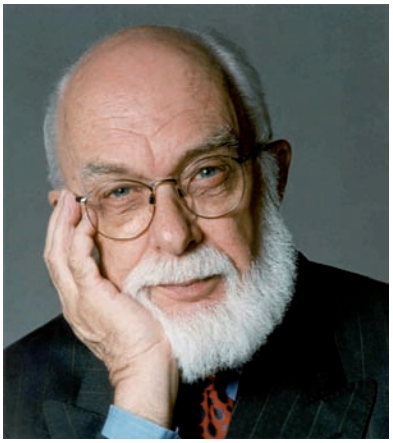


Figure 20 James Randi

Learning checklist

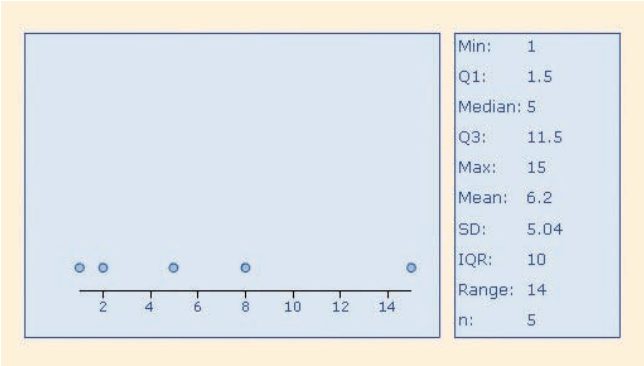
After studying this unit, you should be able to:

- appreciate how various statistical skills and techniques fit into the PCAI stages of a statistical investigation
- create a dotplot of a dataset (using Dataplotter), and understand what it reveals about the shape and location of the data values
- create a boxplot of a dataset (using Dataplotter), and interpret its shape and location
- create a histogram of a dataset (using Dataplotter), and interpret its shape and location
- relate the ideas of location and spread of a dataset (introduced in Unit 4) to pictorial representations of the data in the form of statistical charts
- understand some basic aspects of experimental design
- appreciate the degree of variation that might be expected when items are selected randomly
- appreciate an important idea in statistical decision making, namely, that a useful technique is to compare observed results with results that might be obtained by chance alone.

Solutions and comments on Activities

Activity 1

(a) The screenshot below shows the result of entering the numbers.



(b) The summary values are: Min (the minimum value); Q1 (the lower quartile); Median; Q3 (the upper quartile); Max (the maximum value); Mean; SD (the standard deviation); IQR (the interquartile range); Range; and n (the number of values in the dataset).

(c) The median is a measure of location. If all the data values are sorted by size, from smallest to largest or vice versa, and there is an odd number of data values, then the median is the middle data value. If the number of data values is even, then there is no middle data value, and the median is the mean of the two middle data values in the sorted dataset.

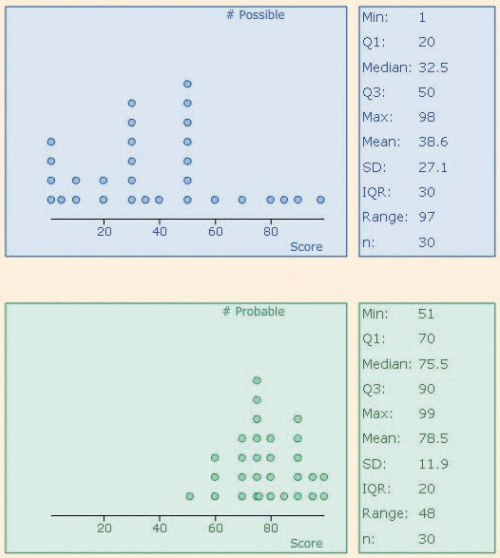
The standard deviation is a measure of spread. Roughly speaking, it gives an idea of how far, on average, the data values are from the mean. It is calculated by finding the deviation (difference) of each data value from the mean of the dataset, squaring each of these deviations, finding the mean of the squared deviations and finally finding the square root of this mean.

Activity 2

(a) Six students gave the value 30.

(b) The two dotplots are shown below.

Dotplots to show the distributions of the 'Possible' and 'Probable' scores
Source: MU123 Book A, Unit 4, Section 3

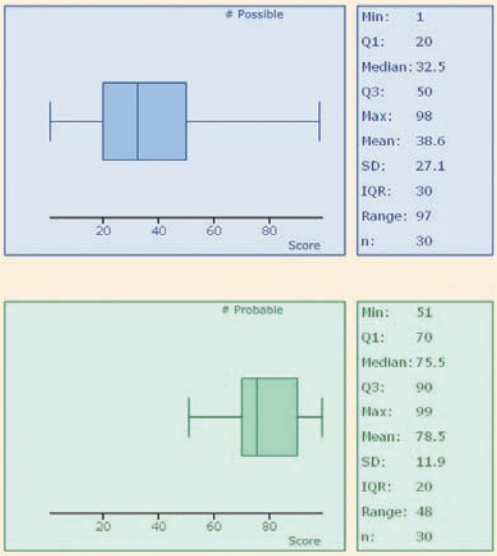


It is clear from the fact that the 'Probable' data are further to the right than the 'Possible' data that the 'Probable' dataset contains generally larger values. Also, the 'Probable' data are more closely packed together – that is, the spread is narrower. These conclusions are in line with what you observed in Unit 4, where you made the comparisons based on numerical summaries of the data.

Activity 3

The two boxplots are shown below.

Boxplots to show the distributions of the 'Possible' and 'Probable' scores
Source: MU123 Book A, Unit 4, Section 3



(b) (i) The median is 75.5.

(ii) The upper quartile is 90.

(iii) The lower quartile is 70.

(c) (i) This statement is correct, because the minimum value of the 'Probable' dataset (which is 51) is greater than the upper quartile of the 'Possible' data (which is 50).

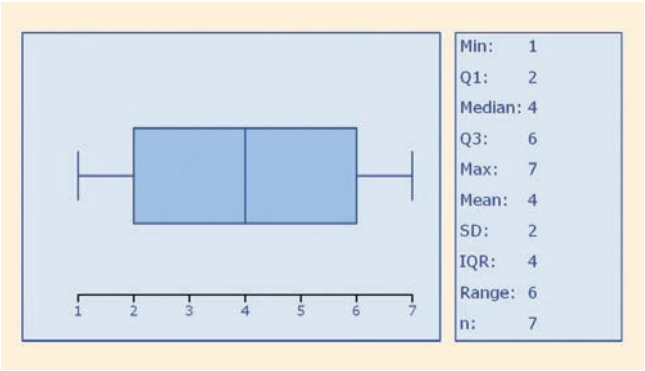
(ii) This statement is incorrect, because the interquartile range of the 'Probable' data is 20, which is *less* than the interquartile range of the 'Possible' data, which is 30.

(iii) This statement is correct, because for the 'Possible' data the difference between the minimum value and the median is $32.5 - 1 = 31.5$, whereas for the 'Probable' data it is $75.5 - 51 = 24.5$.

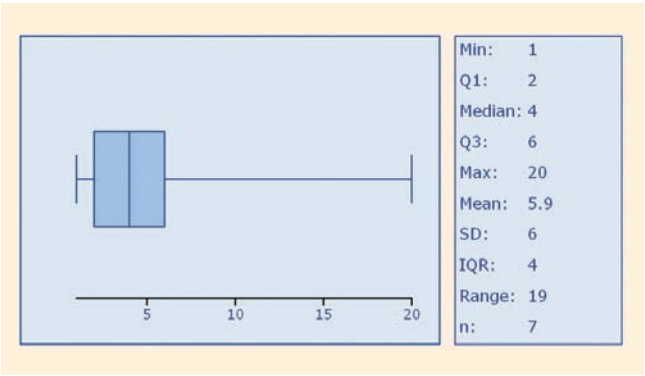
(d) All five key summary values for the 'Probable' dataset are higher than the corresponding values for the 'Possible' dataset; this indicates that the students gave higher numbers for the word 'Probable' overall. Both the length of the box and the length of the boxplot are narrower for the 'Probable' dataset than for the 'Possible' dataset; this indicates that the students were more consistent in the numbers that they gave for 'Probable' than for 'Possible'.

Activity 4

(a) The boxplot is shown below.



(b) When the 7 is changed to 20, the box part of the boxplot appears to shrink and move to the left.



However, this is a misleading observation. In fact, the scale has automatically resized to accommodate the new higher maximum value, and the other four summary values are actually unchanged. They are unchanged because, in this case, the other summary values do not depend on the value of the maximum.

(c) The mean has increased from 4 to 5.9. This has happened because the mean is calculated *from every value in the dataset*.

Activity 5

(a) Each dataset has minimum value 0, maximum value 19 and median 9.5. However, Dataset 1 has lower quartile 7 and upper quartile 12, while Dataset 2 has lower quartile 2 and upper quartile 17, as shown below.

Dataset 1

0	6	7	8	9	10	11	12	13	19
↑		↑		↑			↑		↑
Min		Q1		Median			Q3		Max

Dataset 2

0	1	2	3	9	10	16	17	18	19
↑		↑		↑			↑		↑
Min		Q1		Median			Q3		Max

So Boxplot A represents Dataset 1 and Boxplot B represents Dataset 2.

(Another way that you might have matched up the datasets and boxplots is to observe that the values in the middle half, approximately, of Dataset 2 are more spread out than those in the middle half, approximately, of Dataset 1.)

(b) Each dataset in this part also has minimum value 0, maximum value 19 and median 9.5. However, Dataset 3 has lower quartile 2 and upper quartile 12, while Dataset 2 has lower quartile 7 and upper quartile 17, as shown below.

Dataset 3

0	1	2	3	9	10	11	12	13	19
↑		↑		↑			↑		↑
Min		Q1		Median			Q3		Max

Dataset 4

0	6	7	8	9	10	16	17	18	19
↑		↑		↑			↑		↑
Min		Q1		Median			Q3		Max

So Boxplot C represents Dataset 4 and Boxplot D represents Dataset 3.

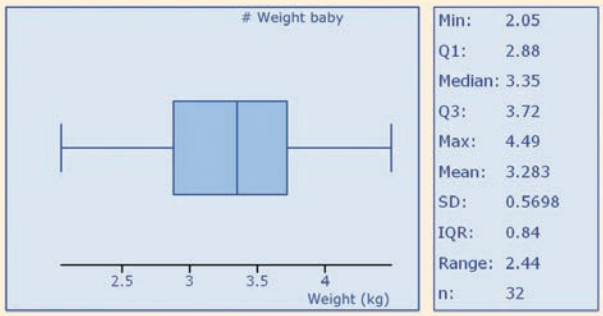
The four correspondences are summarised below.

Dataset	1	2	3	4
Boxplot	A	B	D	C

Activity 6

The boxplot and its summary values are shown below.

Boxplot to show the distribution of the baby weights
Source: MU123 Book A, Unit 4, Subsection 2.1



(a) As expected, the whiskers are longer than the two parts of the box.

(b) This indicates that the middle 50% of the weights are less spread out than the outer 50%. In other words, like most natural phenomena, the baby weights tend to bunch in the middle and are more sparse at the extremes.

Activity 7

(a) The median of the earnings is £400 (from Dataplotter), so the poverty threshold is 60% of £400, which is

$$0.6 \times £400 = £240.$$

(Notice that this is just greater than the value of the lower quartile, £230.)

Three people earn less than £240, so the number of people in poverty is 3.

The percentage of people in poverty is

$$\frac{3}{13} \times 100\% = 23\%$$

(to the nearest whole number).

(b) After the 25% pay rise across the board, the new median is £500, so the new poverty threshold is 60% of £500, which is

$$0.6 \times £500 = £300.$$

(Notice that this is still just greater than the new lower quartile value of £287.50.)

There are still three people earning less than the poverty threshold (those earning £200, £237.50 and £275), so the number of people in poverty is still 3.

Hence the percentage of people in poverty is also unchanged, at 23%.

So the 25% across-the-board rise has made no difference to the number of people in poverty.

(This is the principal reason why the proportion of people in poverty in the UK remained 'stubbornly' at one in five.)

(c) After the £100 pay rise across the board, the new median is again £500, so the new poverty threshold is 60% of £500, which is £300, as in part (b).

(Notice that this is now *less* than the new lower quartile value of £330.)

There are now only two people earning less than the poverty threshold (namely those earning £260 and £290); so the number of people in poverty is 2.

The percentage of people in poverty is

$$\frac{2}{13} \times 100\% = 15\%$$

(to the nearest whole number).

So the £100 across the board rise has had the effect of reducing the number of people in poverty.

Activity 8

(a) There are four data values in the interval 65–70.

(b) When the interval width is 10, there are two peaks, one at 50–60 and a smaller one at 70–80.

When the interval width is 2, there is a high peak at 70–72 and a smaller one at 52–60.

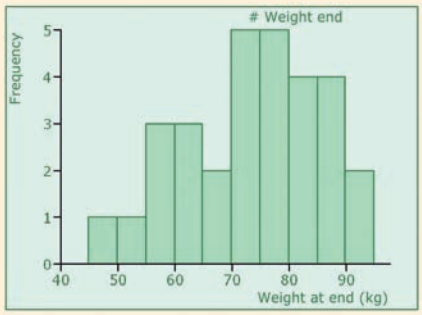
An interval width of 1 gives several peaks. The histogram is now considerably more 'spiky' than those with larger interval widths.

The interval width determines the number of intervals. A small interval width gives a large number of intervals and a spiky histogram.

Activity 9

The 'Weight end' histogram is shown below.

Histogram to show the distribution of the weights of 32 women at the end of their pregnancies
Source: MU123 Book A, Unit 4, Subsection 2.1



(a) In the '# Weight end' dataset, there are more high values than low values.

(Note also the change of scale – when just the 'Weight start' histogram was displayed, the scale on the horizontal axis extended up to 85 kg, but when both histograms are displayed, the scale on both horizontal axes extends up to 95 kg, to accommodate 'Weight end' data values up to 92.7 kg.)

(b) Comparing the two histograms shows that there has been an increase in weights during pregnancy. This is to be expected in healthy pregnancies!

(c) The values of the minimum, the lower quartile, the median, the upper quartile and the maximum are all higher for the 'Weight end' data than for the 'Weight start' data. This indicates that overall the weights increased during pregnancy. Also, both the range and the interquartile range are higher for the 'Weight end' data, indicating that there is more variation in the weights at the end of the pregnancies than in the weights at the start.

It is easier to use the boxplots to compare the datasets, as these charts are simpler and display five summary values. In contrast, the histograms give more details of the distributions.

Activity 10

(a) Histograms A and D each have only one data value in the interval from 0 to 5, so they match with Datasets 1 and 4. Histogram D has only one data value in the interval from 15 to 20, so it matches with Dataset 1, and so Histogram A matches with Dataset 4.

So Histograms B and C match with Datasets 2 and 3. Histogram B has only one data value in the interval from 15 to 20, so it matches with Dataset 3, and so Histogram C matches with Dataset 2.

(An alternative way to match up the datsets and histograms is to look at where the data are concentrated. For example, Histograms C and D are both symmetrical, with concentrations of data at the extremes for Histogram C and in the middle for Histogram D. These clues suggest that these histograms match with Datasets 2 and 1, respectively. Similarly, Histogram A shows a concentration of data in the top interval and Histogram B shows a concentration in the bottom interval, suggesting that these histograms match with Datasets 4 and 3, respectively.)

(b) The datasets match up with the histograms from this activity and the boxplots from Activity 5 as follows:

Dataset	1	2	3	4
Histogram	D	C	B	A
Boxplot	A	B	D	C

As you can see, the boxplots and histograms tell similar stories about the data. With boxplots, regions where the data are concentrated are represented by short whiskers and narrow box sections. On a histogram, these regions show up as having tall columns. For example, if the data values are concentrated in the middle, then the boxplot has a narrow central box and long whiskers (as in Boxplot A on page 14), while the corresponding histogram has tall columns in the middle (as in Histogram D).

General comparisons between histograms and boxplots are given in the main text.

Activity 11

(a) The article quotes the fact that there were three cases of child leukaemia in the Menai Strait area in the period 2000–03, and also uses data for the incidence of child leukaemia nationally in the UK, which gives an expected incidence of 0.1 cases in the Menai Strait area.

(It is not clear whether the article is referring to new cases over the period 2000–03, or all cases known during this period, including pre-existing ones.)

(b) By comparing the expected number of child leukaemia cases, which was close to 0.1, with the actual number of cases, which was 3, the article claimed a 28-fold increase over what was expected. Clearly it is impossible to have 0.1 of a case, so the '28-fold increase' is actually meaningless in this context. A related problem here is that the number of cases is really too small to enable any clear judgement to be made.

Activity 12

Here is a run of numbers that may be fairly typical of the sort of run that you might have written down.

Twenty made-up 'random' numbers:

4 1 9 0 6 3 2 7 8 3 4 6 5 9 3 2 0 8 5 7

Activity 13

Analysing the occurrences of pairs produces the following results.

There are no pairs of consecutive numbers the same in the made-up run.

As shown below, there are two pairs of consecutive numbers the same, and one triple of the same number, in the computer-generated run.

9 0 8 3 7 8 8 6 1 1 1 4 6 6 8 1 8 2 9 2

↑ pair ↑ triple ↑ pair

Activity 14

If you did not avoid repetitions, well done for not imposing orderliness in this way!

Activity 15

A count of the numbers for each run produces the following results.

Number	Frequency in made-up run	Frequency in computer run
0	2	1
1	1	4
2	2	2
3	3	1
4	2	1
5	2	0
6	2	3
7	2	1
8	2	5
9	2	2

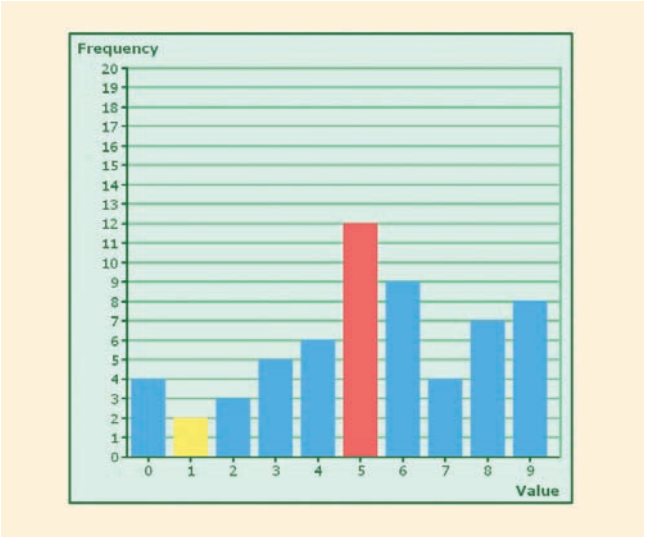
As can be seen from the table, the frequencies of the numbers generated by the computer vary from zero to five, but the frequencies of the made-up numbers vary only from one to three, with eight of the ten numbers having a frequency of two. So, in respect of their frequency of occurrence, the made-up numbers are much more orderly than the computer-generated numbers.

Activity 16

Again, if your frequencies varied quite widely, well done for not imposing orderliness in this way!

Activity 17

(a) The bar chart resulting from a simulation is shown below. (Yours will be different!)



(b) There is considerable variation in the frequencies of the numbers depicted in the bar chart above. The number with the largest frequency is 5 (with a red bar), which occurred twelve times. The number with the smallest frequency is 1 (with a yellow bar), which occurred only twice. So the most frequent number occurred six times as often as the least frequent number.

Activity 18

(a) There is quite a wide variation in the possible ratios, as you'll see in part (b). So the ratio for your run might be substantially higher or lower than the value for the run shown in the solution to Activity 17, which was 6.

(b) The results of one set of ten runs are shown in the table below. (Your results are likely to be rather different.)

Run	Max. freq.	Min. freq.	Max/Min
1	11	3	3.67
2	11	3	3.67
3	11	4	2.75
4	8	3	2.67
5	9	4	2.25
6	13	1	13
7	8	3	2.67
8	11	1	11
9	10	2	5
10	9	3	3

In the ten runs whose results are shown in the table above, the maximum frequencies vary from 8 to 13 (the largest maximum frequency, 13, is marked in red). The minimum frequencies vary from 1 to 4 (the smallest minimum frequency, 1, is marked in yellow). The ratios of maximum frequency to minimum frequency vary from 2.25 to 13.

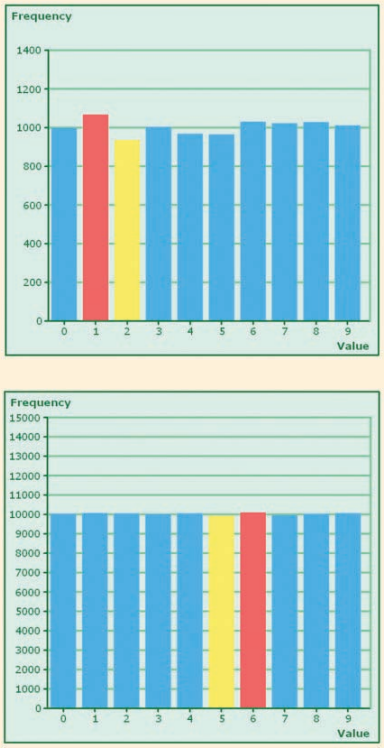
The results of the ten runs given in the table above seem to suggest that the run discussed in the solution to the previous activity (which had a maximum frequency of 12, a minimum frequency of 2, and hence a ratio of maximum frequency to minimum frequency of 6) is by no means atypical, and indeed runs with even greater variation in frequencies may occur.

Activity 19

(a)–(c) You should have observed that usually the more random numbers there are, the less variation there is in the heights of the bars, and the closer the ratio of maximum frequency to minimum frequency is to 1.

This illustrates that, as the number of random numbers is increased, the frequencies of the numbers tend to settle down and become approximately equal.

The bar charts below are fairly typical of what you will have observed for runs of 10 000 and 100 000 random numbers, respectively. In each case, the variation in the heights of the bars is small in comparison to the overall heights of the bars.



Activity 20

(a) On average, the number of cases that each town would expect to have in the year in question is

$$\frac{60}{6} = 10.$$

(b) To use the 'At random' software to generate 60 random numbers between 1 and 6, use the following settings:

Generate values between: 1 and 6

Number of values (per run): 60

Number of runs: 1

Each random number generated represents a case in one of the six towns. For example, an occurrence of the number 2 represents a case in Town 2. So the frequency of each number represents the number of cases in the corresponding town. There are sixty cases altogether, so sixty random numbers need to be generated.

(c) Comments on this part are given in the text.

Activity 21

(a) Cheating would be discovering the number on a card using sensory methods, that is, methods that don't require explanation in terms of ESP.

(b) Your suggestions might have included the following possibilities:

- If the cards look even slightly different from the back (perhaps a little smudged, or of marginally different sizes), then a subject might use that information. So you should ensure that the cards are exactly the same size, and their backs are identical and clean.
- A mirror in the room would obviously not be a good idea! But reflection from the table might also afford a possibility of cheating if it is glass topped. So you should make sure that there are no reflections that could give away the numbers on the cards.
- If you shuffle the cards in the sight of the subject, then someone who really wanted to cheat might be able to track the position of a particular card. So you may consider shuffling the cards behind a screen.
- Your own face may unintentionally give away some information, if you know what the card is at the time that the subject guesses. So you should ask the subject to make their guess before you look at the card yourself.

Activity 23

The screenshot below shows typical results for 100 runs of the simulation. Your dotplot is likely to be different.

Activity 24

(a)–(b) What you are seeing is a series of repeats of the 100 simulations that you carried out in Activity 23, with each set of 100 scores represented by a bar chart rather than a dotplot. You will notice a fair amount of variation in the shape of the bar charts from one set of 100 scores to another. This is due to the fact that the number of scores is small (just 100), and random variation is strongly evident in the results. Typically, you should find that the scores range from around 1 to 11, with a concentration of scores at around 5.

(c) You should find that with 100 000 scores the amount of variation in the shape of the bar charts is much reduced. Also, typically you should find that the scores range from 0 to 13, with a concentration of scores at 5.

Activity 25

(c) You should have obtained the following percentages.

Score	Percentage of scores that are this high or higher (to 1 d.p.)
8	10.2%
9	4.1%
10	1.4%
11	0.4%
12	0.1%

(d) By the results of part (c), a score needs to be at least 9 to fall into the top 5% of scores. That is, the minimum score that would be regarded as showing evidence of ESP is 9.

(e) By the results of part (c), a score needs to be at least 11 to fall into the top 1% of scores. That is, the minimum score that would be regarded as showing evidence of ESP in this case is 11.

(The answers in parts (d) and (e) are the same as the answers based on just 100 scores on page 43, but this is just by luck! Some other sets of 100 scores would have given different results.)

Activity 26

(a) (i) In Activity 25(d) it was found that the 5% critical value is 9.

The solution to Activity 25(c) shows that approximately 4.1% of scores obtained simply by guessing will be 9 or more; that is, will lie in the 5% critical region.

So you would expect about 4.1% of subjects taking the test to achieve a score in the 5% critical region purely by chance.

(ii) The number of subjects taking the card-guessing test in a given presentation of MU123 is roughly

$$2000 \times 2 \times 0.90 = 3600.$$

So you would expect the number of subjects who achieve a score in the 5% critical region purely by chance to be approximately

$$3600 \times \frac{4.1}{100} \approx 148.$$

(b) (i) In Activity 25(e) it was found that the 1% critical value is 11.

The solution to Activity 25(c) shows that approximately 0.4% of scores obtained simply by guessing will be 11 or more; that is, will lie in the 1% critical region.

So you would expect about 0.4% of subjects taking the test to achieve a score in the 1% critical region purely by chance.

(ii) From part (a), the number of subjects taking the card-guessing test in a given presentation of MU123 is roughly 3600, so you would expect the number of subjects who achieve a score in the 1% critical region purely by chance to be approximately

$$3600 \times \frac{0.4}{100} \approx 14.$$

(You could check your answers to parts (a)(ii) and (b)(ii) by using the 'By chance' tab of the 'By chance alone' software to generate 3600 scores typical of those that might be obtained by guessing. You can do this by clicking 'Reset' and ticking 'Cumulate data', then generating three lots of 1000 scores followed by six lots of 100 scores. You can use the bar chart (or the percentage obtained by using 'Shade score') to see roughly how many of these scores are greater than or equal to 9, or greater than or equal to 11.)

(c) Probably neither the 5% criterion nor the 1% criterion is strict enough, as it is likely that with either of these criteria a reasonable number of subjects will achieve, purely by chance, a score good enough to be considered as evidence of ESP.

Activity 28

A sensible way to proceed with a subject who achieved a high score would be to ask him or her to take the test again several times. If he or she consistently maintained high scores, then and only then might you be looking for an explanation! You would want to consider explanations such as cheating, and whether you had carried out the experiment correctly, before concluding that your subject did indeed show evidence of ESP.